# DOCUMENT RESUME

| ED 051 116 | 24 | SP 004 994 |
|---|---|---|

| | |
|---|---|
| AUTHOR | Rosenshine, Barak |
| TITLE | Interpretive Study of Teaching Behaviors Related to Student Achievement. Final Report. |
| INSTITUTION | Temple Univ., Philadelphia, Pa. Coll. of Education. |
| SPONS AGENCY | National Center for Educational Research and Development (DHEW/CE), Washington, D.C. |
| BUREAU NO | BR-9-B-010 |
| PUB DATE | Sep 70 |
| GRANT | OEG-2-9-480010-1016(010) |
| NOTE | 163p. |
| EDRS PRICE | EDRS Price MF-$0.65 HC-$6.58 |
| DESCRIPTORS | *Academic Achievement, *Classroom Observation Techniques, *Educational Research, Interaction, *Research Methodology, Research Needs, Research Problems, *Teacher Behavior |

ABSTRACT

This report reviews the results of some 35 studies completed between 1956 and 1970 which attempted to relate systematically observed teaching behaviors to adjusted measures of student achievement. Information in each study includes investigator, date, population, time, tests used, and significant and non-significant results. The studies are divided into four categories according to the type of behavior investigated: 1) affective variables, 2) teacher cognitive behaviors, 3) flexibility and variety, and 4) amount of teacher-student interaction. In the first category, consistent positive trends were noted for use of student ideas, indirectness, and indirect/direct ratios, and a consistent negative trend for criticism. There are too few studies in the second category for any generalizations, but in the third category, variation in activities was positively related to student achievement. In the fourth category, there were consistently positive but non-significant correlations between teacher talk and student achievement. Suggestions for further research include the use of a greater variety of variables, the use of high and low inference variables in the same investigation, subdivision of variables, greater control over the relationship between instructional content and criterion measures, and greater precision in recording, reporting, and analyzing results. (RT)

ED051116

Final Report
Project No. 9-B-010
Grant No. OEG-2-9-480010-1016(010)

Approved
John Sokol 11/23/70

Dr. John Sokol
Director, Educational Research
DHEW-OFFICE OF EDUCATION - RM 1013
Federal Building
26 Federal Plaza
New York, New York 10007

# INTERPRETATIVE STUDY OF TEACHING BEHAVIORS RELATED TO STUDENT ACHIEVEMENT

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Barak Rosenshine

College of Education
Temple University
Philadelphia, Pa. 19122

September, 1970

The research reported herein is the Final Report of Project Number 9B-010 from the Small Grants Research Projects of the Office of Education, U.S. Department of Health, Education, and Welfare. The contract was awarded to Temple University.

United States Department of Health, Education, and Welfare

National Center for Educational Research and Development

SP004994

# ACKNOWLEDGEMENT

This report was written under the continuous help and supervision of N.L. Gage, Norma Furst, Barbara Rosenshine, Robin Nelson, Robert Soar, Ted Amidon, Francis Sobol, Graham Nuthall, Ned Flanders, Richard Snow, A.W. Foshay, and Frances Martin. Additional contributions were made by investigators who supplied me with original data from their study and who answered my letters promptly. The assistance of Rebecca Guth of the Temple Instructional Materials Center in arranging for the purchase and dissertations, use of the Temple University ERIC collection, and use of the microfiche reader-printer was invaluable. A great deal of this report was written on the ranch of Vera and Raymond Lewelling in St. Helena, California. The setting and their hospitality provided inestimable help.

Many of the ideas in this report could not have been developed without the daily conversations with Norma Furst.

## Table of Contents

Chapter One - Introduction

Selection of Studies 1.2

Independent Variables 1.2

Dependent Measures 1.3

Number of Studies 1.4

Limitations in Comparing Studies and Using Results 1.4

1 & 2. Developing and Using an Observational Instrument 1.5

3. Determining Student Achievement 1.6

Determining Significant Relationships 1.7

Correlation and Causation 1.10

Limitations of Results 1.10

Explanation of Tables 1.13

Interaction Analysis 1.14

Chapter Two - Affective Variables

Investigators' Source of Variables 2.1

Criticism and Control 2.2

Teacher Non-Verbal Approval 2.13

Teacher Use of Praise 2.13

Use of Student's Ideas 2.20

Combined or Unique Measures of Teacher Approval 2.25

Ratio of Teacher Approval to Teacher Disapproval Statements 2.28

Summary 2.35

Table of Contents (Continued)

Chapter Three - Teacher Cognitive Behaviors

Types of Questions 3.4

Teacher Questions--Classified Into Two Types 3.4

Types of Questions--Multiple Classification 3.11

Other Cognitive Variables 3.17

Probing 3.17

Structuring 3.19

Task Oriented Behavior 3.24

Clarity 3.27

Summary 3.28

Chapter Four - Flexibility and Variety

Summary 4.8

Chapter Five - Amount of Teacher-Student Interaction

Teacher Talk 5.1

Student Talk 5.4

Teacher-Student Interaction 5.6

Summary 5.9

Chapter Six - Future Research on Teaching Behaviors

Summary of Results 6.1

Future Process-Product Research 6.3

Selection of Variables 6.4

Procedures for Coding Classroom Events 6.7

Design of Process-Product Studies 6.13

Methods of Analysis 6.15

Reporting Results 6.21

Experimental Classroom Studies 6.22

Conclusion 6.23

Table of Contents (Continued)

Appendix

1. Clarity A.2
2. Variability A.4
3. Enthusiasm A.5
4. Task Oriented and/or Businesslike A.6
5. Student Opportunity to Learn Criterion Material A.7
6. Use of Student Ideas and General Indirectness A.8
7. Criticism A.10
8. Use of Structuring Comments A.11
9. Types of Questions A.12
10. Probing A.13
11. Level of Difficulty of Instruction A.14
Summary of Process-Product Results A.15
References

## Table of Tables

| Table | | Page |
|---|---|---|
| 1.1 | Definitions of the Independent Variables | 1.16 |
| 2.1 | Teacher Use of Criticism or Disapproval | 2.3 |
| 2.2 | Non-Verbal Approval | 2.14 |
| 2.3 | Teacher Praise | 2.16 |
| 2.4 | Acceptance of Student Ideas | 2.23 |
| 2.5 | Combined or Unique Measures of Teacher Approval | 2.26 |
| 2.6 | Ratios of Teacher Approval Statements to Teacher Disapproval Statements (i/d Ratios) | 2.30 |
| 3.1 | Types of Teacher Questions: Two Classifications | 3.6 |
| 3.2 | Multiple Classification of Questions or Interactions | 3.14 |
| 3.3 | Probing | 3.18 |
| 3.4 | Structuring | 3.20 |
| 3.5 | Task Oriented, Achievement Oriented, Businesslike | 3.25 |
| 4.1 | Flexibility | 4.2 |
| 4.2 | Variety and Variation | 4.5 |
| 5.1 | Teacher Talk | 5.2 |
| 5.2 | Student Talk | 5.5 |
| 5.3 | Teacher-Pupil Interactions (Frequency of Questions) | 5.7 |

# Chapter One

## Introduction

The purpose of this chapter is to review the results of some 35 studies which have attempted to relate systematically observed teaching behaviors to adjusted measures of student achievement.

Fifteen years ago, two reviews were written about the relationship of teacher characteristics and behaviors with measures of student achievement (Morsh and Wilder, 1954; Ackerman, 1954). In 24 of the 25 studies reported by those reviewers, the teacher characteristics included presage variables such as age, intelligence, experience, or scores of the teacher on personality tests; and the teacher behaviors were assessed by rating scales marked by students, supervisors, or the investigator. The reviewers concluded that the results of the studies were contradictory and inconsistent, and recommended the use of systematic observation techniques in future studies of teacher behaviors which may be related to pupil achievement:

> Because the actual behavior of the teacher in the classroom is such an important factor, it is necessary to devise means of observing and recording this behavior. Methods must be used in which only a minimum of inference is allowed.... Such a process does suggest a potentially wider range of investigation which it is hoped will provide more reliable information in the areas of teacher effectiveness and pupil change (Ackerman, 1954, pp. 286-287).

Proposals of this sort were well received in the educational community, and soon many workers were developing objective, reliable, observational category systems which did not rate but counted the frequencies of specified teacher behaviors. Eighty different category systems for classroom observation have been collected in the 15 volume anthology, Mirrors for Behavior (Simon and Boyer, 1970); at least eighty others could be uncovered with little effort.

It is relatively easy to develop observational systems and obtain high interrater agreement. The majority of investigators have used their systems to describe teaching (e.g., to report that classrooms have a certain proportion of teacher talk, pupil talk, divergent questions, use of pupil ideas, or evaluation by public private criteria). But at some point we should ask which of the hundreds of behaviors that can be objectively and reliably counted are related to pupil growth. Many of these behaviors ought to have significant correlations with pupil growth, but as Gage has noted, "We have been fooled before in educational research and I, for one, shall rest uneasy until the evidence on these plausible but undemonstrated connections is in" (Gage, 1966, p. 35).

Some evidence is in. This review focusses upon investigations in which category systems have been used for something more than description and in which attempts have been made to determine specific relationships between what a teacher does and what pupils learn. It is offered as a sympathetic review.

## Selection of Studies

The major studies in this review are those in which the investigators used the natural setting to find relationships between specific teaching behaviors and pupil achievement. All these studies are labeled correlational, although a number of investigators used an F-test or a t-test to determine the level of significance of their findings.

### Independent Variables

In order to sharpen our focus, the only studies selected for this review were those in which classroom observational category systems were used to code specific teacher behaviors. Following the initial recommendations of Ackerman (1954), recent reviewers (Gage, 1969; Rosenshine, 1970) referred to such variables as "low inference measures" because the items in the observational category systems focus upon specific, denotable, relatively objective behaviors such as "teacher repetition of student ideas," or "teacher asks evaluative question," and because these events are recorded as frequency counts.

Classroom observational rating systems have been classified as "high inference measures" because they lack such specificity. Items on rating instruments such as "clarity of presentation," "enthusiasm," or "helpful towards students" require that an observer infer these constructs from a series of events. In addition, the observer must also infer the frequency of such behavior in order to record whether it occurred "consistently," "sometimes," or "never," according to whatever set of gradations is used in the scale of an observational instrument. To a reader, the statement that a teacher repeated student answers 7% of the time is much more specific than the statement that a teacher was sometimes helpful towards students. Gage (1969) has noted that it is difficult to translate such dimensions as "responsive," "clear," or "achievement oriented" into specific ways of behaving.

In this review, only studies which employed low inference measures are included. The results on all observational studies of teaching (i.e., those which used observational category systems, observer rating scales on specific behaviors, and student questionnaires) are currently being brought together in a single volume (Rosenshine, in preparation). The major significant results to date have been summarized (Rosenshine and Furst, in press) and are presented as an appendix to this report.

## Dependent Measures

In order to focus this review further, measures of student achievement were the only dependent measures considered. Other criterion measures (which were also studied in some of the investigations reviewed below) include student interest, student liking for teacher, amount of homework turned in, type of student questions, amount of level of student participation, or work oriented behavior, but they were not considered in this review. These measures were excluded because the strength and consistency of the correlation between these measures and residual class mean achievement scores has not been adequately established. In addition, many of these measures appear to be of sufficient educational concer to merita separate review on the relationship between specific teacher behaviors and student growth in these areas. Results on the relationships between teaching behaviors and other criterion variables such as student creativity, student anxiety, or student attitudes towards school and school subjects were also excluded because these areas also appear to merit a separate review.

## Number of Studies

Approximately 35 studies are included in this review. The precise number cannot be specified because several studies were not completely independent of each other. For example, Harris et al. (1968) studied the same teachers across two years, but the students were different in each analysis. Should this be counted as one or two studies? In the study by Powell (1968), the same students were studied across two years, but the students had different teachers. In the two studies reported by Wallen (1966), the two samples of teachers and students were independent, but they were observed by the same raters. Because of this overlap, and because different reviewers might classify the number of studies reported by the above investigators differently, no more precise term than "approximately 35 studies" can be applied.

Some studies were excluded from this review because the number of teachers was less than 10, or because residual achievement scores were not obtained. Other studies may have been omitted because the reviewer was unaware of them. I would greatly appreciate any reference to additional process-product studies.

## Limitations in Comparing Studies and Using Results

Usually the conducting of this type of research includes four steps: (1) develop an instrument which can be used systematically to record the frequency of certain, specified teacher behaviors, (2) use the instrument to record the classroom behaviors of teachers and their pupils, (3) rank the classrooms according to a measure of pupil achievement adjusted for the initial difference among the classes, and (4) determine the behaviors whose frequency of occurrence is related to the adjusted class achievement scores.

As might be expected in a new area of study, the investigators differed widely in the procedures which they used in order to complete each of the above four steps. The variety of procedures makes it most difficult to compare and synthesize the studies in this area. In addition, there are unresolved methodological problems at each of these four steps which further complicate the comparison of the studies, the evaluation of the results, and the strength of any recommendation for the use of these findings in teacher training

(Cf. Biddle, 1967; Meux, 1967; Rosenshine and Furst, in press; Rosenshine, 1970c). Some of the difficulties most relevant to the reading of this review and the independent evaluation of my synthesis are elaborated below. The difficulties are considered under each of the four steps.

## 1 & 2. Developing and Using an Observational Instrument

The observational instruments described in this review contain a set of non-evaluative, relatively objective categories to describe what goes on in the classroom. In the process of developing inter-rater agreement, each investigator had to develop many ground rules to clarify distinctions between such items as "questions about content" and "questions that stimulate thinking" (Perkins, 1964), and "praise and encouragement" and "use of pupil ideas" (Flanders, 1965). Some reports included detailed descriptions of the coding protocols (e.g., Bellack et al., 1966; Spaulding, 1965), whereas others provided only the names or short definitions of the observed behaviors. The reports that did not include detailed protocols require the reader to make more than a minimum of inference in interpreting the results of the studies. For example, although nine investigators employed the Interaction Analysis system (Flanders, 1965) for recording classroom behavior, only one investigator (Snider, 1966) specified the ground rules used to distinguish between different categories. Although each investigator reported high inter-rater reliability, the degree of inter-investigation reliability remains uncertain: we do not know whether raters trained by Flanders (1965), Soar (1966), or Furst (1967) would have agreed in their scoring of behaviors if they all viewed the same classroom. The possibility of systematic differences between investigators who are using the same category system can be empirically tested by having a number of investigators who use the same category system (e.g., Flanders' Interaction Analysis) code the same set of audiotapes or videotapes and determining the amount of agreement or disagreement between them. Although such studies have been proposed, I have not read of the results of any such study.

The possibility of systematic differences between different investigators is increased when categories are developed which are relatively ambiguous (or high-inference). For example, the category in the Flanders system labeled "teacher use of student ideas" has been subjected to different interpretation by different investigators.

Even the same investigators have differed in the exact definition of "low inference" variables. In the first edition of the booklet providing a complete description of Flanders' Interaction Analysis system, Flanders and Amidon (1963) wrote that teacher statements which are repetitions of a student's words are coded as "Category 3." In a later edition (Flanders and Amidon, 1967), the same behavior was placed in "Category 2" (praise). Later, Flanders (1970) wrote that repeating the main words a student said is a subcategory of "Category 3" (Use of student ideas).

Because of the possibility of systematic differences between observers at different institutions, the low inference systems described in this review might best be called relatively objective observational instruments. Therefore, it is difficult to compare the results of studies by different investigators, and even more difficult to suggest the specific behaviors which might be taught in a teacher training program. This problem of conceptual clarity might be overcome; at present, it appears to an artifact of this relatively new approach to the study of classroom instruction.

## 3. Determining Student Achievement

The second phase of any process-product study is ranking the classrooms on some measure of achievement. There are at least two problems in interpreting the results in this area: the method of computing the measures of student gain, and the comparability of different achievement tests.

Residual Gain Measures. In all the studies selected for inclusion in this review, regression procedures were used to adjust the posttest scores for measures of initial achievement and/or aptitude. Although the adjustment procedures were usually labeled analysis of covariance or residual gain scores, the specific procedures differed from study to study. The investigators also differed in the variables which they used as covariates. Some used a single subject area pretest; some used multiple subject area covariates; some used measures of learning aptitude; some used achievement and aptitude covariates. The extent to which different statistical procedures would have yielded different results is a topic of recent discussion (Cf. Coats, 1966; Lord, 1962; Wallen and Wodtke, 1963). Indeed, the appropriateness of any residual gain procedure in situations in which random assignment was not possible and in which systematic differences may exist on variables other than the covariate (s) has been questioned (Cronbach and Furby, 1970).

Different Criterion Instruments. The problems of computing residual gain scores aside, some studies have been conducted in which seemingly similar criterion instruments have yielded different measures of class mean residual gain, so that the correlations between teacher behavior and student achievement were different with different instruments. For example, Snider (1966) used both the New York State Regents Exam and the Cooperative Physics Test as his criterion instruments, yet teacher behaviors which were related to residual gain measures on one instrument were not significantly related on the other. In a study of teacher ratings, Chall and Feldman (1966) found that teacher ratings which were significantly related to student achievement in reading on the Stanford Achievement Test were not significantly related to scores on the Fry Reading Test or the Gates Reading Test. Finally, teacher behaviors which were significantly related to reading achievement but not arithmetic achievement in one study, showed the opposite result in another study.

As this review was written, I found myself referring to "significant results" if a significant correlation was obtained on one or two of four criterion instruments. Such a slip appears natural when one is attempting to find teaching behaviors that yield consistent correlations with student achievement. There remains the possibility that results which were significant when one criterion instrument was used, might not have been significant if another instrument had been used, and *vice versa*.

## Determining Significant Relationships[1]

The statistical procedures which the investigators used to relate teacher behaviors and student achievement were varied and are not easily compared. In general, three types of statistical procedures were used: Simple correlation, inferential statistics, and factor analysis.

---

[1] In *all* instances in which the word "significant" is used in this report, the term is taken to mean results which were statistically significant at the .05 level of confidence or better. No other meaning of the term is used of implied. In this report, "significant findings" are limited to statistically significant findings; no implication of *educational* significance is intended unless the educational value of the findings is discussed in the text. The reader is encouraged to reinterpret the results and to suggest their educational relevance.

Simple Correlation. Simple correlation was the most frequently used statistical technique and was employed in 20 studies. Almost all the investigators computed product-moment correlations, although one used rank order correlations (Cook, 1967) and one investigator computed a tau (LaShier, 1967). Few of the investigators computed more than 10 correlations. In some cases, however, a large number of correlations was computed in an effort to explore a variety of hypotheses. For example, Wright and Nuthall (1970) computed 37 simple correlations between measures of teacher behavior and student achievement, and six of these correlations were significant at the .05 level. Unfortunately, we do not know which two of these might have occurred by chance.

The interpretation of factor loadings created difficulty in assessing whether or not significant results were obtained. For example, in the study by Soar (1966), the variable "teacher nonverbal affection" had a loading of .56 on a factor which had a significant correlation of .30 with residual gain in vocabulary. Because of the size of the loading and the factor correlation, "teacher nonverbal affection" cannot be said to be a significant variable by itself; yet, it cannot be labeled a non-significant variable because it loaded on a significant factor. In the summary of each set of results in Chapters Two through Five, variables which loaded on a significant factor were included as representing significant results, although neither the factor scores nor the factor loadings was included when the range of rs was given in the body of the text.

In some cases a variety of affective and cognitive variables loaded on the same factor so that the same factor seemed relevant under a number of classifications. In developing the integrative tables below, factors were included under a specific variable if it contained loadings for variables which appeared relevant to the particular table. In order to conserve space, whenever a factor appeared on a number of tables, only those variables relevant to the particular table (or set of variables under discussion) were presented. In order to integrate the factor analytic studies with the others, they were classified according to variables, and each factor was repeated on every table which focused on a variable on that factor. For example the same factor in the study by Spaulding (1965) appears in the tables on disapproval (Table 1.1), praise (Table 1.3), and task oriented

(Table 3.2). In each instance, only those variables relevant to the variable under consideration are presented, instead of repeating all the loadings of .40 or above on that factor. The appropriate factor loadings are given in each table, and these loadings are correlations with the factor score, not correlations with achievement.

Variation in statistical procedures. The variety of statistical procedures used makes comparison and synthesis of the results extremely difficult, and makes any conclusions hazardous. The fact that an investigator reported results as statistically significant does not mean that he would have obtained significant results if he had used other analytic procedures. Similarly, non-significant results might have been significant if other analytic procedures had been used.

One estimate of the power of different statistical procedures might be obtained by examining the results of five studies in which both correlational and inferential statistics were used to analyze the results. In all the studies, the level of significance was drastically reduced when a correlation was computed, and statistically significant results remained in only one study. In that study, a probability of .001 which was obtained when extremes of teachers were selected and students were used as the sampling unit (Morrison, 1966) was reduced to a probability of .05 when a correlation was computed using all teachers (Flanders, 6th grade, 1970). In the other studies, results which were significant at the .01 level when an F or t was computed, were not significant at the .10 level when an r was computed (Furst, 1967; Soar, 1966; Flanders, 8th grade, 1965 and 1970). In one instance a Critical Ratio reported at a probability greater than .001 when students were the sampling unit, became a correlation of .48 ($p > .10$) when a correlation was computed using class means. These results suggest that it is easier to obtain statistically significant results when inferential statistics are used.

Statistical significance. This reviewer believes that statistical significance in itself is not a sufficient criterion for accepting or rejecting the possibility of a relationship between a teacher behavior and student achievement. A correlation coefficient of .20 is educationally meaningless no matter how many asterisks follow the coefficient. But a series of studies on the same variable, all yielding positive correlations of .20 can be indicative of a consistent relationship which is worthy of attention whether or not the coefficients are statistically significant. The reader is supplied with all the information which this reviewer used to make his decisions, and is encouraged to inspect the data in this review and in the original studies and reach alternative conclusions according to his purposes.

## Correlation and Causation

Different purposes dictate different designs. If the purpose of the investigation is to differentiate between high-, middle-, and low-achieving teachers so that this information can be used in subsequent experimental studies, then inferential designs would seem most appropriate. At the same time, the reader should note that this review, in the tradition of studies of teacher effectiveness, does not focus on studies in which teachers have been trained to exhibit certain behaviors. Rather, the focus is upon those studies in which naturally occurring teacher behaviors have been related to measures of student achievement. The results of such correlational studies should not be taken as indicators of causation.

Because of the variety of statistical procedures used in these studies, a common term, "significant relationship," was used to describe all significant results regardless of the statistical procedure which was used.

## Limitations of Results

Given the problems in developing and using observational category systems, in calculating student gain, and in relating observed behaviors to student achievement, and conclusions reached in this review must be seen as extremely tentative.

When the proposal for this review was written, it was hoped that there would be sufficient consistency in the results to allow some of the best findings to be used in teacher education programs. Currently, such a hope appears to be beyond the available data. Perhaps the best we can do at present is to view the most promising variables in this review as hypotheses for future experimental studies. In such a framework, questions of design, inter-investigator reliability, and statistical procedures become less crucial. It now appears that the best use of these results is not in training teachers to behave in certain ways; rather, the best hope may be in designing and conducting experimental studies to determine whether training teachers in the most promising variables can result in enhanced student achievement.

The last line of the box describing the study contains the length of time between pretest and posttest, or the length of time of the instructional period. Within the text, studies in which the instructional period was one hour or less are frequently described as "short term" studies. Studies whose instructional period was one semester or longer were described as "long term" studies. A "semester" is approximately five months long. Studies which were conducted across a school year were identified as lasting "two semesters." A term such as "two semesters" is only an approximation because investigators differed in the time they selected for administering the pretest. In many cases the investigator (s) administered the pretest during the first or second month of the school year. But some investigators chose to use as pretests the standardized achievement tests which had been administered at the end of the previous school year.

The identical descriptive left-hand box was used each time different variables from a study were discussed within the following chapters. For example, the study by Soar (1966) appears in Chapters two, Three, Four, and Five and appears in more than one place within these chapters. Therefore, these tables were developed to provide relevant information each time different aspects of the study were mentioned.

Middle and Right-side Columns. The second and third columns contain the significant and non-significant results for each study. "Significant" refers to statistical significance at the .05 level of confidence or better. Results which are significant at the .05 level are indicated by a single asterisk (*). Following the usual conventions, two asterisks (**) refer to significance at the .01 level or better, and three asterisks (***) refer to the .001 level. Results significant at the .01 level are marked with the footnote (a).

Most of the investigators used correlational statistics to relate teacher behavior and student achievement; however, some investigators used inferential statistics such as analysis of variance (or covariance) or a t-ratio. In the tables, the type of statistic which was used is identified in the first line of the cell describing each of the significant and non-significant results. In summarizing the results of these studies, the term "relationship" is sometimes used, even though some of the investigators employed inferential statistics. Some investigators used both correlational and inferential statistics, and whenever possible, this reviewer reanalyzed studies in which only inferential statistics were used in order to present the results obtained using both procedures.

One study (Hunter, 1968) was completely reanalyzed by this reviewer in order to provide data on correlational procedures using the class as the statistical unit.

In some investigations, five or six criterion variables were used, and only one or two were significant. In those cases, the significant results and the tests were presented in the center column. When all tests in the battery yielded significant results, the median correlation was presented in the middle column. This median correlation was expressed by the abbreviation "med." or "mdn." If significant results were obtained on one of five criterion tests, then the single significant correlation was presented in the middle column, and the median correlation for the five tests was presented in the right-hand column under "non-significant results." These procedures were adopted in order to relieve the reader of the burden of reading even longer lists of results.

Most of the investigators who used an F-test used a one-factor, two-level analysis of variance (or covariance) procedure. In those few cases where the investigator split his sample into three groups, this fact is indicated by the words "trichotomized sample."

## Explanation of Tables

The tables in Chapters Two, Three, Four, and Five represent an effort to compress a great deal of information into readable form. They represent the best solution which the reviewer and his advisers found for the problem of presenting the reader with complete, yet manageable information.

Left-hand Box. The major identification of each study is in the box in the left hand column. The first line (s) give the investigator and the year of publication. The next line gives the grade level (s) of the student. The conventions followed in the U.S.A. are used to identify the grade levels. First grade students are usually six years old;eighth grade students are thirteen years old.

The grade level is followed by the major subject area covered in the criterion instrument (s). The term "General" was used whenever a battery of achievement tests covering a large number of subject areas was administered. The specific tests used in each investigation are also given in the second column of the identification tables at the end of this introduction.

The number of teachers in each study is presented in parenthesis. A notation such as (15 tchrs) indicates that there were 15 teachers in the sample and all teachers were used in the analysis. A notation such as (16/55 tchrs) indicates that of the original sample of 55 teachers, 16 teachers who were either high-achieving or low-achieving were selected for analysis. Whenever possible, further descriptive information was included in the text.

## Interaction Analysis

Because eight of the studies discussed here have used Interaction Analysis (IA) to describe teacher and pupil behavior and two others have used modifications of the IA categories, it is necessary to describe the categories, the use of the matrix, and the development of variables from the cells of the matrix.

IA refers to the systematic observational procedures developed by Flanders (1965). (Cf. Amidon and Flanders, 1967.) All verbal classroom behavior is coded into one of ten categories:

1. Teacher accepts student feelings
2. Teacher praises or encourages student
3. Teacher accepts or uses student idea
4. Teacher asks question
5. Teacher lectures
6. Teacher give directions
7. Teacher criticizes student or justifies authority
8. Student predictable response
9. Student initiated response
10. Silence or confusion

Every three seconds (or more often if the behaviors change more frequently) the observer notes which category best describes the ongoing behavior. The result is a record of classroom behavior expressed in a two-dimensional 10 X 10 matrix which is developed by pairing each category number in the sequence with the number that follows it. Frequencies in specific cells refer to the number of times one behavior followed another (see Figure 1). For example, entries in the cell formed by row 4 and column 8 (area K in Figure 1) refers to the number of times a teacher question was followed by a predictable pupil response. Tallies in the 3-3 cell indicate the extended use of a pupil's idea, or three seconds of repeating or elaborating a pupil idea followed by additional repetition or elaboration.

After the matrix has been constructed, investigators use various combinations of some of the 100 cells to develop variables descriptive of types of teaching behaviors. Coats (1966) described twenty-seven variables developed from the matrix, and at least twenty more have been developed by others.

This large number of variables has resulted in some confusion when different investigators applied the same label to different combinations of cells, or labeled the same combinations of cells with different titles. For example the terms i/d and Revised I/D refer to the identical combination of cells (Table 1.1). The reader of a research report should be careful to check the operational definitions given by the investigator and should not assume that all investigators use the same variables when they refer to a "direct" or an "indirect" teacher.

The operational definitions of some of the common IA variables which will be discussed in this review are presented in Table 1.1.

Figure 1

Selected Interaction Analysis Variables

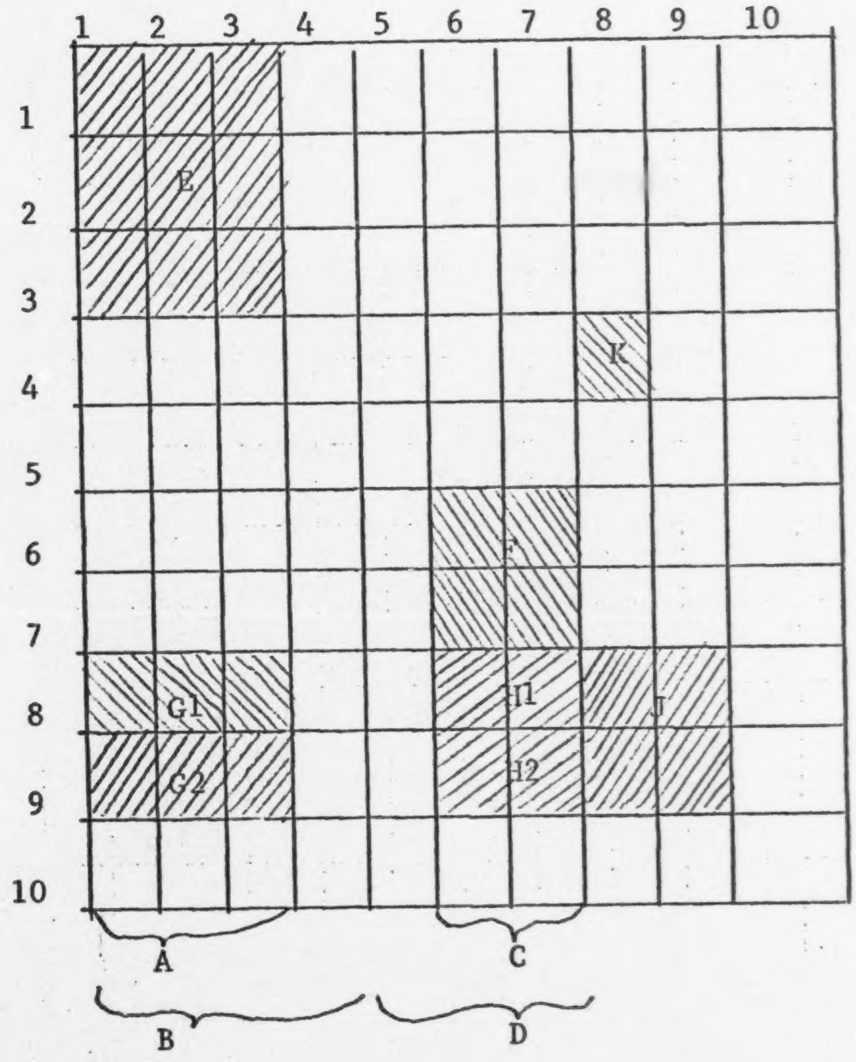
1
2
3
4
5
6
7
8
9
10
E
K
F
G1
G2
H1
H2
J
A
B
C
D

Table 1.1

Definitions of the Independent Variables

| Name | Definitions |
|---|---|
| I/D ratio | Ratio of the number of tallies in columns 1-4 to the number of tallies in columns 5-7 (ratio of area B to area C in Figure 1). |
| i/d ratio | Ratio of the number of tallies in columns 1-3 to the number of tallies in columns 6 and 7 (ratio of area A to area C in Figure 1). |
| i/d 8 | The i/d ratio only for row 8 (ratio of area $G_1$ to are $H_1$). |
| i/d 9 | The i/d ratio only for row 9 (ratio of area $G_2$ to area $H_2$ ). |
| i/d 8-9 | The i/d ratio for rows 8 and 9 (ratio of area $G_1 + G_2$ to area $H_1 + H_2$ ). |
| extended indirect | Percentage of tallies in the following cells: 1-1, 1-2, 1-3, 2-1, 2-2, 2-3, 3-1, 3-2, 3-3 (area E). |
| extended direct | Percentage of tallies in the following cells: 6-6, 6-7, 7-6, 7-7 (area F). |
| extended i/d ratio | Ratio of the number of extended indirect tallied to the number of extended direct tallied (ratio of area E to area F). |

TABLE 1.2

Complete List of Studies and Test Instruments

| STUDY | POSTTEST |
|---|---|
| Anthony, 1967<br>5th - General<br>(21 tchrs)<br>One semester | Stanford Achievement Tests (Average Score). |
| Birkin, 1967<br>5th - Reading<br>(34 tchrs)<br>20 weeks | Silent Reading Tests |
| Conners and Eisenberg, 1966<br>(38 tchrs)<br>6 weeks | Peabody Picture Vocabulary Test |
| Cook, 1967<br>10th - Biology<br>(8 tchrs)<br>Two semesters | Iowa Tests of Educational Development<br>Watson-Glaser Critical Thinking Appraisal<br>Processes of Science Test<br>BSCS Comprehensive Final Exam |
| Flanders, 1970<br>2nd - General<br>(15 tchrs)<br>Two semesters | Stanford Achievement Tests (Mean Score)<br>Two semesters |

| STUDY | POSTTEST |
|---|---|
| Flanders, 1970<br>4th - Social Stud.<br>(16 tchrs)<br>Two weeks | Special social-studies unit<br>Two weeks |
| Flanders, 1970<br>6th - General<br>(30 tchrs)<br>Two semesters | Metropolitan Achievement Tests (Mean Score)<br>Two semesters |
| Flanders, 1970<br>7th - Social Stud.<br>(15 tchrs)<br>Two weeks | Special social-studies unit<br>Two weeks |
| Flanders, 1970<br>8th - Math<br>(16 tchrs)<br>Two weeks | Special math unit<br>Two weeks |
| Flanders (1965)<br>7th - Social Stud.<br>(15 tchrs)<br>Two weeks | Special Social studies Unit |
| Flanders (1965)<br>8th - Math<br>(16 tchrs)<br>Two weeks | Special Math Unit |
| Furst, 1967<br>10th and 12th grades - Social Studies<br>(15 tchrs)<br>Four one-hour lessons | Special tests |

| STUDY | POSTTEST |
|---|---|
| Harris and Serwer, 1966<br>1st - Reading<br>(48 tchrs)<br>Two semesters | Stanford Achievement Tests. (Separate scores for word reading, para. meaning, vocabulary, spelling word study skills)<br>Two semesters |
| Harris *et al.*, 1968<br>2nd - Reading<br>(38 tchrs)<br>Two semesters | Metropolitan Achievement Tests (Separate scores for word knowledge, word discrimination, reading, and spelling).<br>Two semesters |
| Hunter, 1968<br>Educationally Handicapped Children,<br>ages 8 to 14<br>(11 tchrs)<br>Two semesters | Wide Range Achievement Test |
| Kleinman, 1964<br>7th, 8th - Science<br>(6 of 23 tchrs)<br>Cross-sectional study | Test on Understanding Science |
| LaShier, 1967<br>8th - Biology<br>(10 tchrs)<br>Six weeks | BSCS Unit Test |
| Medley and Mitzel 1959.<br>3rd thru 6th - Reading<br>(49 tchrs)<br>Two semesters | California Reading Test |

| STUDY | POSTTEST |
|---|---|
| Morsh, 1956<br>Airmen (Mechanics)<br>(109 tchrs)<br>Seven one-hour sessions | Special test in aircraft hydraulics |
| Perkins, 1965<br>5th - General<br>(27 tchrs)<br>Two semesters | California Achievement Tests (Separate scores for language arts, reading, social studies, and arithmetic) |
| Penny, 1969<br>8th and 9th - Social Studies and English<br>(32 tchrs)<br>Two 45-minute sessions | Special tests for each lesson |
| Powell, 1968<br>3rd - Reading and Arithmetic<br>(9 tchrs)<br>Two semesters | Science Research Associates Tests in Reading and Arithmetic |
| Powell, 1969<br>4th - Reading and Arithmetic<br>(17 tchrs)<br>Two semesters | Science Research Associates Tests in Reading and Arithmetic |
| Schirner, 1968<br>High School - Earth Sciences<br>(17 tchrs)<br>Two semesters | Test on Understanding Science<br>Test of Science Knowledge, Pts. I and II<br>Earth Science Curriculum Project Final<br>Earth Science Final |

| STUDY | POSTTEST |
|---|---|
| Sharp, 1966<br>High School - Biology<br>(31 tchrs)<br>Two semesters | Nelson Biology Test |
| Shutes, 1969<br>8th and 9th - Social Studies and English<br>(32 tchrs)<br>Two 45-minute sessions | Special tests for each lesson |
| Snider, R. M., 1966<br>12th - Physics<br>(17 tchrs)<br>Two semesters | New York Regents Physics Exam<br>Cooperative Physics Tests<br>Test on Understanding Science |
| Soar, 1966<br>3rd thru 6th - General<br>(55 tchrs)<br>Two semesters | Iowa Tests of Basic Skills (Separate scores for reading, vocabulary, arithmetic problems, and arithmetic concepts) |
| Solomon _et_ _al_., 1963<br>College evening school<br>American History<br>(24 tchrs)<br>One semester | Special tests on facts and comprehension in American History |

| STUDY | POSTTEST |
|---|---|
| Spaulding, 1965<br>4th and 6th Reading and Mathematics<br>(21 tchrs)<br>Two semesters | Sequential Tests of Educational Progress (Separate scores for reading and mathematics) |
| Thompson and Bowers, 1968<br>4th - Vocab. and Social studies<br>(15 tchrs)<br>Two semesters | Stanford Achievement Tests (Separate scores on vocabulary and social studies) |
| Torrance and Parent, 1966 (1st study)<br>7th thru 12th<br>SMSG-Math<br>(33 of 75 tchrs)<br>Two semesters | STEP - Math |
| Vorreyer, 1965<br>5th, General<br>(14 tchrs)<br>Two semesters | California Achievement Test (Separate scores for vocabulary, reading, language arts, arithmetic, and social studies) |
| Wallen, 1966<br>1st - General<br>(36 tchrs)<br>Two semesters | California Achievement Tests. (Separate scores on vocabulary, reading, and arithmetic) |
| Wallen, 1966<br>3rd - General<br>(40 tchrs)<br>Two semesters | California Achievement Tests. (Separate tests on vocabulary, reading, and arithmetic) |

| STUDY | POSTTEST |
|---|---|
| Wallen and Wodtke, 1963<br>1st thru 5th - General<br>(65 tchrs)<br>Two semesters | California Achievement Tests (Separate scores for vocabulary, reading, and arithmetic) |
| Wright and Nuthall 1970<br>3rd, Science<br>(17 tchrs)<br>Three ten-minute lessons | Special test on science materials |

## Chapter II

## Affective Variables

This section attempts to synthesize the results of 30 studies in which frequencies of teacher affective behaviors were related to measures of student achievement. The affective variables were divided into six categories: criticism and control, non-verbal approval, praise, use of student ideas, indirectness, and variables which represent a ratio of approving and disapproving behaviors. Each table includes only those studies which appeared to include variables in one of the six categories, and each table presents the results across all achievement criterion measures used in each investigation. These divisions are tentative and should be revised as the results of future investigations are reported.

Each part of this section begins with a description of the variable, contains a summary of the findings, and closes with recommendations for future study. Throughout the report, the word "significant" refers to statistical significance at the .05 level or better.

The tables on specific approval or disapproval behaviors are limited in that they include only those studies which provided information on the relationships between these variables and achievement. Some investigators who used IA or OScAR (Medley and Mitzel, 1959) as their observational instrument also counted instances of approval and disapproval, but their independent variables were some combination of these behaviors into an i/d ratio or a measure of "supportiveness." A reanalysis of the original data or IA matrices to isolate frequencies of specific approval and disapproval behaviors may clarify the relationships between those variables and achievement, and the results summarized below may be changed when such reanalysis is completed.

### Investigators' Source of Variables

Almost every investigator has included an affective variable such as approval or disapproval in his study of correlates of cognitive achievement. This choice is well founded in experimental research: generations of psychologists have studied the effects of positive and negative reinforcement upon learning, and textbooks in educational psychology include sections on the results of this research. Travers (1967) listed 66 generalizations which he considered to be the most significant in terms of their applicability to understanding and guiding classroom practice, and almost a quarter of these generalizations are on positive and negative reinforcement. Investigators specializing in behavior modification have begun systematically to apply these variables in situations similar to the natural classroom (Orme, 1968; Wasik et al., 1968; Gallagher and Aschner, 1967; O'Leary and Becker, 1967).

But although most observation systems contain affective categories, the authors seldom cite the above research as justification for including praise or criticism among their categories. Instead, as Wallen and Travers (1963) and McDonald (1963) have noted, the authors refer to philosophical positions or to a line of research beginning with H. H. Anderson (1939) or Levin, Lippitt, and White (1939). References to Skinner are absent from the reviews of research.

## Criticism and Control

Seventeen studies were found which included variables which might be labeled "teacher criticism of students" (Table 2.1). In most of the studies linear correlations were computed between different measures of criticism and pupil achievement in various subjects, but in four studies the investigators used factor analytic techniques (Anthony, 1967; Perkins, 1965; Soar, 1966; Spaulding, 1965), and linear correlations are not available for these four studies.

A single table describing the results of 17 studies is too gross a summary because a variety of behaviors, ranging from giving simple directions to extreme teacher hostility are contained in these variables. The specific categories which one investigator developed overlap those another developed, and so this table cannot be divided easily into smaller tables. However, an attempt is made to describe

Table 2.1

Teacher Use of Criticism or Disapproval (Counting)

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Anthony, 1967<br>5th, General<br>(21 teachers)<br>One semester | number of instances of observed negative affect<br>(one of 14 items on total scale. Scale r with ach. = .48) | |
| Cook, 1967<br>10th, Biology<br>(8 teachers)<br>Two semesters | | Criticism (Column 7), mdn. rho = -.33<br>Extended criticism (Cell 7-7), mdn. rho = -.33 |
| Flanders, 1970<br>2nd, General<br>(15 teachers)<br>Two semesters | | IA<br>Col. 6&7 (teacher direct behavior), r = -.10<br>Cells 6-7 and 7-6 (extended criticism), r = .05<br>Restrictive feedback, r = .18 |
| Flanders, 1970<br>4th, Social Stud.<br>(16 teachers)<br>Two weeks | | IA<br>Col. 6&7, r = -.24<br>Cells 6-7 & 7-6, r = -.23<br>Restrictive feedback, r = -.34 |

Table 2.1 (cont.) Teacher Use of Criticism

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Flanders, 1970<br>6th, General<br>(30 teachers)<br>Two semesters | | IA<br>Col. 6&7, *r* = -.04<br>Cells 6-7 & 7-6, *r* = -.15<br>Restrictive feedback,<br>r = -.32 |
| Flanders, 1970<br>7th, Social Stud.<br>(15 teachers)<br>Two weeks | IA<br>Col. 6&7, *r* = -.61*<br>Cells 6-7 & 7-6, r + -.62*<br>Restrictive feedback,<br>*r* = -.50 | |
| Flanders, 1970<br>8th, Math<br>(16 teachers)<br>Two weeks | | Col. 6&7, *r* = -.34<br>Cells 6-7 & 7-6, *r* = -.24<br>Restrictive feedback,<br>*r* = -.43 |
| Harris and<br>Serwer, 1966<br>1st, Reading<br>(48 teachers)<br>Two semesters | negative motivation,<br>*r* with spelling = .29*<br><br>teacher control,<br>med. *r* = .28* | negative motivation,<br>med. *r* = .16 (all *r*'s<br>were positive) |
| Harris et al.,<br>1968<br>2nd, Reading<br>(38 teachers)<br>Two semesters | negative motivation,<br>*r* with reading = -.40* | negative motivation<br>med. *r* = -.26 (all 4<br>*r*'s were negative)<br><br>teacher control,<br>med. *r* = -.19 |

Table 2.1 (cont.) Teacher Use of Criticism

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Hunter, 1968<br>Emotionally handicapped children<br>ages 8 to 14,<br>General<br>Two semesters | hostile or strong disapproval, med. $r$ = -.61* | directive statements related to school, med. $r$ = -.23<br>neutral or mild disapproval, med. $r$ = -.21<br>teacher justification of authority, med. $r$ = -.38 |
| Morsh, 1956<br>Airmen<br>(Mechanics)<br>(109 teachers)<br>seven hours | | teacher gives directions, $r$ = .10<br>Teacher threatens or warns, $r$ = .05 |
| Perkins, 1965<br>5th, General<br>(27 teachers)<br>Two semesters | Factor II, Teacher Lecturer-Criticizer | |
| | teacher criticizes | teacher criticizes |
| | +[a] reading vocabulary<br>- reading comprehension<br>- English grammar | ns[a] arithmetic reasoning<br>ns arithmetic fundamentals<br>ns spelling |
| | Factor III, Teacher Leading Recitation | |
| | teacher rejects or corrects student response | |
| | + arithmetic reasoning | ns reading<br>ns arithmetic fundamentals<br>ns English grammar<br>ns spelling |

---

[a] + or - refers to positive or negative loading on a factor containing this behavior. "ns" refers to no loading on a factor containing this variable.

Table 2.1 (cont.) Teacher Use of Criticism

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Perkins, 1965, cont. | Factor IV, Student Individual Work | |
| | teacher rejects or corrects student response | |
| | + spelling | ns reading<br>ns arithmetic<br>ns English |
| | teacher gives directions or commands (did not load on any of the four factors) | |
| Soar, 1966<br>3rd thru 6th,<br>General<br>(55 teachers)<br>Two semesters | Factor 1, Teacher Criticism | |
| | Pupil initiation following teacher criticism (-.74)[a]<br>Teacher verbal hostility (-.76)<br>Continued teacher criticism (Cell 7-7) (-.83) | |
| | *r* = .29* (arith. concepts)<br>*r* = .34* (arith. problems) | *r* = .16 (vocabulary<br>*r* = .13 (reading) |
| | Factor 5, Unnamed[b] | |
| | Continued criticism and directions (IA Cells 6-6 & 6-7 & 7-6 & 7-7) (-.84) | |
| | | mdn. *r* = .03 |

---

[a] Only those component variables related to criticism are given here; all the loadings on each component are not given. Loading directions have been reflected, when necessary to show negative relationships.

[b] The coefficients in parentheses refer to component loadings; they *do not* represent correlations with the criterion measures. The *r* represents the correlation between the *total component* and the criterion measure.

Table 2.1 (cont.) Teacher Use of Criticism

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Spaulding, 1965<br>4th and 6th,<br>reading and<br>mathematics<br>(21 teachers)<br>Two semesters | Component 2[a] | |
| | total disapproval (-.42)[b]<br>disapproval by veiled or explicit threat to do harm (-.44) | |
| | $r$ = .49* (reading) | $r$ = .10 (math.) |
| | Component 6 | |
| | disapproval by commanding conformance (.41)<br>disapproval by eliciting clarification in a non-threatening way (.36) | |
| | $r$ = .44* (reading) | $r$ = .39 (math.) |
| | Component 10 | |
| | disapproval by social shaming or sarcasm (-.55)<br>disapproval by anonymous or impersonal warnings (-.44) | |
| | $r$ = .42[c] (reading) | $r$ = .08 (math.) |

---

[a] Only those component variables related to criticism are given here; all the loadings on each component are not given. Loading directions have been reflected, when necessary to show negative relationships.

[b] The coefficients in parentheses refer to component loadings; they *do not* represent correlation with the criterion measures. The $r$ represents the correlation between the *total component* and the criterion measure.

[c] $p < .10$

Table 2.1 (cont.) Teacher Use of Criticism

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Wallen, 1966<br>1st, General<br>(36 teachers)<br>Two semesters | personal control<br>*r* = -.38* (vocab.) | personal control<br>*r* = -.22 (arithmetic)<br>*r* = -.08 (reading<br>(comprehension)<br>academic control<br>*r* = ns (specific *r*'s not reported) |
| Wallen, 1966<br>3rd, General<br>(40 teachers)<br>Two semesters | | personal control<br>med. *r* = -.22<br>academic control (ns) |
| Wright and Nuthall, 1970<br>3rd, Science<br>(17 teachers)<br>Three 10-minute lessons | | teacher managerial comment, *r* = -.22<br>challenging comment<br>*r* = -.38 |

clusters of behaviors within the larger variable "criticism and control," but the definitions which investigators gave may not be comparable, and these definitions may not be identical to the operational definitions which the observers developed in the course of coding.

Results. Of the 17 studies, one showed significant negative linear correlations (Flanders, 7th grade, 1970), seven yielded significant negative relationships on at least one criterion measure (Anthony, 1967; Harris et al., 1968; Hunter, 1968; Perkins, 1965; Soar, 1966; Spaulding, 1965; Wallen, 1st grade, 1966), one showed significant positive relationships on at least one criterion measure (Harris and Serwer, 1966), and in eight studies non-significant relationships were obtained (Cook, 1967; Flanders, 2nd grade, 1970; Flanders, 4th grade, 1970; Flanders, 6th grade, 1970; Flanders, 8th grade, 1970; Morsh, 1956; Wallen, 3rd grade, 1966; Wright and Nuthall, 1970). In other words, significant negative correlations between teacher use of criticism and student achievement on at least one criterion measure were obtained in half of the 17 studies.

If only the direction of the correlation is considered, negative correlations between any measure of criticism and all measures of student achievement were obtained in 12 of the 17 studies, and these correlations ranged from -.04 to -.62. Positive correlations for all variables were obtained in two studies (Harris and Serwer, 1966; Morsh, 1956), but these correlations tended to be small ($r$'s from .05 to .29). Mixed results were obtained in two studies (Perkins, 1965; Spaulding, 1965) and will be discussed below.

Mild Criticism. Several investigators developed categories of mild forms of criticism or control, such as the giving of academic directions. In no study did mild criticism have a significant negative relationship with achievement. Thus, Hunter (1966) did not find significant correlations for "neutral or mild disapproval" or for "directive statements related to school;" Perkins (1965) did not find that giving directions loaded on any factors; Spaulding (1965) did not find that disapproval by negative evaluation loaded on a significant factor; and Wallen (1966) did not find significant correlations between academic control and student achievement.

In two studies, mild criticism was positively related to achievement. Perkins found that the behavior "teacher does not accept student's answer" loaded on the same factor as the total class gain in arithmetic, and Spaulding found that disapproval both by commanding conformance and by eliciting clarification in a non-threatening way loaded on a factor positively related to achievement in reading.

The four investigators who found that mild criticism was not related to achievement or was sometimes positively related to achievement also found that strong criticism had significant, negative relationships with achievement. Hunter found significant results for "hostile or strong disapproval" (med. $r$ = -.61); Perkins found that criticism loaded on the same factor as achievement measures; Spaulding found that "total disapproval," and "disapproval by shaming or threat" loaded on significant factors; and Wallen found significant results for personal control in Grade 1.

Affect Loading of Criticism. In 16 of the studies (Anthony, 1968, is excluded) it is possible to compare the relationship of different types or intensities of criticism to pupil achievement. In ten of these studies the stronger form of criticism had a higher negative correlation with achievement than the milder form(s). Thus, in three of the five studies by Flanders (1970), teacher criticism or directions following a student statement had a higher negative correlation with achievement than the sum of teacher use of criticism and teacher giving of directions (Flanders, 1970, Grades 4, 6, and 8). Harris et al. (1968) found that "negative motivation," or teacher statements intended to make the student feel bad, yielded a significant negative correlation with reading ($r$ = -.40), whereas teacher statements designed to control the class yielded smaller and non-significant correlations. Hunter (1968) modified the category system developed by Withall (1961) so that there were separate classifications for "hostile or strong disapproval" and "neutral or mild disapproval." Hostile disapproval yielded a significant negative correlation ($r$ = -.61), whereas mild disapproval had a correlation coefficient of -.21. Perkins (1965) found that teacher criticism had a negative loading on the same factor as total class growth in both reading and English, whereas the variable "teacher does not accept student answer" had a positive loading on the same factor as total class growth in arithmetic reasoning. Soar (1966) found that continued teacher criticism had a negative loading on a factor which was significantly related to student achievement in arithmetic concepts (r = .29) and arithmetic problems (r = .34), but a mixture of giving directions and criticism did not load on a significant factor. Spaulding (1965) found "disapproval by veiled or explicit threat to do harm," "disapproval by social shaming," and "disapproval by impersonal warnings" all had negative loadings on factors significantly related to growth in reading (factor r's = .49 and .42), whereas "disapproval by eliciting clarification in a non-threatening way" had a positive loading on a factor which was significantly related to growth in reading (r = .44). In two studies by Wallen (included in a single report, 1966) the variable "personal control" yielded higher negative correlations with achievement than did the variable "academic control." Academic control refers to the teacher directing the student to perform certain actions clearly related to academic learning; personal control refers to statements directed towards the students' personal rather than academic behavior. Finally, in the study by Wright and Nuthall (1970) teacher challenging comments yielded a higher negative correlation with student achievement (r = -.38) than teacher managerial comments (r = -.22), although neither correlation was significant.

These distinctions between the affect loading for forms of criticism appear useful, but it should be noted that the distinctions were clear in only 10 of the 16 studies. The review of research appears to indicate there is no evidence to support a claim that a teacher should avoid telling a student that he is wrong, or should avoid giving academic directions. However, teachers who use a great deal of criticism appear consistently to have classes who achieve less in most subject areas.

Strong disapproval and criticism was a significant correlate not only in studies of disadvantaged children (Harris et al., 1968) but also in studies involving upper middle class students (Perkins, 1965), upper middle class students with above average ability with teachers rated as superior (Spaulding, 1965), and teachers who were comparatively highly indirect (Soar, 1966). Soar (p. 189) developed a table to show that the teachers in his sample had higher i/d ratios than those in the samples studied by Flanders (1965) and Furst (1967); yet Soar found that teacher criticism was a significant correlate. In the study by Spaulding (1965) ten percent of the mean teachers' behavior was classified as overtly disapproving, compared with twelve percent approving behavior; yet the disapproving behavior had the greatest effect.

One puzzling finding obtained by Perkins (1965) was that teacher criticism loaded on the same factor as total class gain in reading vocabulary. The remaining results on this factor were as expected: teacher criticism was related to total class loss in reading comprehension and mechanics of English. This finding is in the opposite direction from the trends and significant findings in all the other studies.

In the study by Spaulding (1965), the technique of disapproval appeared to be more important than the topic which was disapproved. Thus, disapproval by threat, shaming, and warning was negatavely related to reading achievement, whereas disapproval by commanding conformance and disapproval by eliciting clarification in a non-threatening manner were positively related.

Discussion and Recommendations. The existing correlational research on teacher disapproval or teacher criticism appears inadequate because insufficient attention has been given to the context in which these behaviors occur. In the studies above, only Spaulding (1965) developed a category system which specified the teacher tone, technique, topic, and basis for disapproval, and the results to such subdivision were most useful (as reported above). It is recommended that in future studies the affect loading of the criticism or disapproval, the events preceding and following the disapproval, and the content or event being criticized all be examined. Items referring to teacher disapproval should also be separated from items referring to teacher approval and not subsumed under a general category such as "teacher warmth." Such a recommendation is made because teacher approval statements and teacher disapproval statements were not significantly correlated in the only two studies for which such data were available (Soar, 1966; Hunter, 1968).

Little research has been done on the relationship between teacher, student, and observer perceptions of teacher disapproval or approval. An event which is noted as reflecting disapproval when seen by an observer may not have the same meaning to a student, and vice versa.

It is extremely important that some attempt be made to determine whether any relationship exists between teacher disapproval and cognitive aspects of the teacher's behavior. There has been almost no research in this area. There are suggestions from the research of Solomon et al, (1963), and Wright and Nuthall (1970) that some aspects of teacher criticism may occur when the teacher is unclear, and the class responds by asking for clarification, but studies which included detailed cognitive and affective teacher behavior were rare.

## Teacher Non-Verbal Approval

Only four investigations were found in which teacher non-verbal affective behaviors were counted (Morsh, 1956; Soar, 1966; Wallen, 1st grade, 1966; Wallen, 3rd grade, 1966) (Table 2.2), and in no study was there a clear correlation between teacher non-verbal affection and a measure of student achievement. Counts of teacher non-verbal affection did load positively on one of the strongest and most significant factors in the study by Soar (med. $r$ = .28), but this variable was the only teacher behavior to load on the factor. The other loadings on this factor were for student verbal hostility (-.66) and a rating of student interest and attention (.65). Furthermore, teacher non-verbal affection did not have significant zero order correlations with any of the achievement measures. That this factor was the strongest correlate of overall achievement and yet was almost without teacher behaviors is a surprising and disappointing finding.

Because of the lack of research in this area there are inadequate data for making any generalization as to the importance of teacher non-verbal approval.

## Teacher Use of Praise

The results of 15 studies in which teacher use of praise was counted (Table 2.3) are not as consistent or as strong as those obtained in the review of teacher use of criticism or disapproval. The results are difficult to summarize because of the variations in design. In three studies, more than one criterion measure was used, and the results are different for

Table 2.2

Non-Verbal Approval (Counting)

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Morsh et al., 1955<br>Airmen (Mechanics)<br>(106 teachers)<br>seven hours | | teacher smiles or laughs, $r$ = -.09 |
| Soar, 1966<br>3rd thru 6th, General<br>(55 teachers)<br>Two semesters | Factor 6, Teacher Support<br>teacher non-verbal affection (.56)<br>med. $r$ = .28*[a] | |
| Wallen, 1966<br>1st, General<br>(36 teachers)<br>Two semesters | | teacher non-verbal affection<br>ns (correlations not given in complete report) |
| Wallen, 1966<br>3rd, General<br>(40 teachers)<br>Two semesters | | teacher non-verbal affection<br>ns (correlation not given) |

---

[a] All other loadings on this factor referred to student behaviors.

different criterion measures (Spaulding, 1965; Wallen, 1st grade, 1966; Wallen and Wodtke, 1963). In these same three studies, different forms of praise or approval yielded different results. In two studies (Anthony, 1967; Spaulding, 1965) the statistical significance of positive correlations between measures of praise and student achievement cannot be determined because these variables are presented as loadings on significant factors. In two studies (Perkins, 1965; Wallen, 3rd grade, 1966) the direction of the non-significant correlations was not given in the final report.

Significant positive correlations (or loadings on significant factors) relating some aspect of teacher praise to at least one criterion measure were obtained in 5 of the 15 studies. Positive and significant linear correlations were obtained in three studies (Flanders, 6th grade, 1970; Wallen, 1st grade, 1966; Wright and Nuthall, 1970) and positive factor loadings in two studies (Anthony, 1968; Spaulding, 1965). Nine studies showed non-significant results. Significant negative relationships between praise and achievement were obtained in one study (Wallen and Wodtke, 1963) but were not replicated in a subsequent study (Wallen, 1966).

Discussion. Although there is a tendency toward a positive relationship between teacher approval and pupil achievement, the directions of the correlations are inconsistent from one study to the next. These inconsistent results suggest that approval is such a gross variable that the context, source, type, and topic of approval should be considered.

Some findings have interesting implications for future research. For example, Wallen found that although praise was not a significant correlate for first grade students, both minimum reinforcement and the frequency of the teacher's asking questions had positive correlations with the adjusted achievement scores. Minimum reinforcement was defined as positive reinforcement which is less strong than praise, e.g., "Uh huh," "Right," "Okay." This combination suggests that for the first grade, practice rather than encouragement is the significant variable. However, the observation system developed by Wallen did not include tallying of student behavior, and so this suggestion cannot be studied using his data.

Table 2.3

Teacher Praise (Counting)

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Anthony, 1967<br>5th, General<br>(21 teachers)<br>One semester | Items in summed composite[a]<br>Instances of observed teacher positive support<br>Number of observed achievement awards in room<br>$\underline{r}$ = .48* (for summed composite with achievement) | |
| Flanders, 1970<br>2nd, General<br>(15 teachers)<br>Two semesters | | Teacher use of praise (Col. 2), $\underline{r}$ = .25 |
| Flanders, 1970<br>4th, Social Stud.<br>(16 teachers)<br>Two weeks | | Teacher use of praise (Col. 2), $\underline{r}$ = -.15 |
| Flanders, 1970<br>6th, General<br>(30 teachers)<br>Two semesters | Teacher use of praise (Col. 2), $\underline{r}$ = .36* | |

---

[a] Of the 14 variables in the composite, only those relevant to praise are included here.

Table 2.3 (cont.) Praise

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Flanders, 1970<br>7th, Social Stud.<br>(15 teachers)<br>Two weeks | | Teacher use of praise (Col. 2), $r$ = -.23 |
| Flanders, 1970<br>8th, Math.<br>(16 teachers)<br>Two weeks | | Teacher use of praise (Col. 2), $r$ = -.30 |
| Harris and Serwer, 1966<br>1st, Reading<br>(48 teachers)<br>Two semesters | | positive motivation, med. $r$ = .14<br>(all five $r$'s were positive) |
| Harris et al., 1968<br>2nd, Reading<br>(38 teachers)<br>Two semesters | | positive motivation, med. $r$ = -.19<br>(all 4 $r$'s were negative) |
| Hunter, 1968<br>Emotionally handicapped children, ages 8 to 14,<br>General<br>(11 teachers)<br>Two semesters | | teacher praise or elaboration of student idea, median $r$ = .46<br>(all 3 $r$'s were positive) |

Table 2.3 (cont.) Praise

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Perkins, 1965<br>5th, General<br>(27 teachers)<br>Two semesters | | teacher praise and encouragement, non-significant for all 5 criterion measures |
| Spaulding, 1965<br>4th and 6th, reading and math.<br>(21 teachers)<br>Two semesters | Component 6[a]<br>approval regarding student's interpretation (.52)[b]<br>$r = .49$* | |
| | Component 10[a]<br>approval source: teacher-centered I (-.66)[b]<br>approval source: appeal to convention (.73)<br>approval regarding pupil planning (.44)<br>$r = .42$[c] (reading) | $r = .08$ (math) |
| | Component 12[a]<br>total approval (.51) | $r = .15$ (reading)<br>$r = .08$ (math.) |

---

[a] The name of the component or factor is not given because only those variables specific to the category under consideration are presented in this table.

[b] Refers to component or factor loading. This coefficient is not a correlation with any achievement measure.

[c] $p < .10$

Table 2.3 (cont.) Praise

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Wallen and Wodtke, 1963<br>1st thru 5th, General<br>(65 teachers)<br>Two semesters | teacher demonstrates affection<br>med. $r$ = -.61** (sic)<br>(for arith. gain)<br>(correlations at each grade were negative) | teacher demonstrates affection<br>no significant correlations at any grade reported for vocabulary or reading comprehension |
| Wallen, 1966<br>1st, General<br>(36 teachers)<br>Two semesters | minimum reinforcement (e.g. "uh huh," "Okay")<br>med. $r$ = .39* | praise and encouragement ns on three criterion measures (coefficients not reported)<br>recognizes pupil's raised hand, ns on three criterion measures (coefficients not reported) |
| Wallen, 1966<br>3rd, General<br>(40 teachers)<br>Two semesters | | minimum reinforcement, ns on three criterion measures (coefficients not reported)<br>praise and encouragement ns (same as above)<br>recognized pupil's raised hand, ns (same as above) |
| Wright and Nuthall, 1970<br>3rd, Science<br>(17 teachers)<br>Three 10-minute lessons | teacher gives thanks and praise<br>$r$ = .49* | |

The research of Spaulding suggests that the topic of praise may be more important than the frequency. In his investigation total instances of approval did not load on a significant component (Table 2.3). However, there were positive loadings for two topics of approval: approval regarding student's interpretation, and approval regarding student's planning. Other topics of approval--personal qualities, accurate knowledge, attention to task, and personal interests--did not load on a significant component. Approval regarding student's interpretation and student's planning would appear to be critical for developing cognitive independence and appropriate for the above average students whom Spaulding studied (the sample mean was the 86th centile on the School and College Ability Test). Different topics of approval may be important for students of low ability, and there may be interactions between the type of approval and the cognitive styles of the students. These questions remain to be investigated.

The research by Spaulding also suggests that not all approval is related to achievement. Approval through "teacher-centered 'I'," the use of a warm voice, and the selection of instructional topics related to the pupils' interests all appeared to be negatively related to achievement.

It is unfortunate that those investigators who used IA did not inspect the correlation of cell frequencies with achievement. One interesting variable might be extended praise (Cell 2-2) because such praise contains a reason for the praise; another might be praise in response to student-initiated questions (Cell 9-2).

In sum, research of this type has not shown that there is a consistent linear relationship between the frequency of approval and achievement, and, therefore, the question of whether curvilinear relationships exist remains open. However, the research does suggest that certain types and topics of approval may be positively related to achievement, and that some forms of approval may be negatively related to achievement.

## Use of Student's Ideas

Another form of approval is "teacher accepts or uses ideas of pupil" (Flanders, 1965). Behaviors in this area, coded as Category 3 in the Interaction Analysis scheme developed by Flanders, include the following (Flanders, 1970):

1. Acknowledges the pupil's idea by repeating the nouns and logical connectives he has expressed.

2. Modifying the idea by rephrasing it or conceptualizing it in the teacher's own words.

3. Applying the idea by using it to reach an inference or take the next step in a logical analysis of a problem.

4. Comparing the idea by drawing a relationship between it and ideas expressed earlier by a pupil or the teacher.

5. Summarizing what was said by an individual pupil or a group of pupils.

Behaviors in Flanders' Category 3 would appear to be more powerful affective variables than praise for two reasons: First, repetition of, summary of, and referral to students' ideas seem to be related to two of the greatest tributes in the academic world: being published and being cited. Second, a teacher does not necessarily have to listen to a student in order to give praise: a perfunctory, "Very good," can be given at random moments, or can be used to end a rambling statement by the student to which the teacher does not wish to devote attention. But a teacher must listen and engage in implicit practice in order to apply, compare, summarize, or even repeat an idea. Therefore, the use of students' ideas may be a more intensive form of praise than saying, "Fine," or "Very good."

Not only is Category 3 potentially important as a positive reinforcer, but it also may be an important cognitive variable, in providing repetition, summary, or illustration.

Because of the importance of Category 3 on an intuitive basis, it is unfortunate that we have little specific information on the effects of using students' ideas. Most investigators who have used the IA matrix have included all or part of column 3 as part of an i/d ratio, but few have studied the effects of this variable alone.

Results. Nine studies were found which considered the teacher's use of student ideas, and not one yielded a significant linear correlation between the use of this variable and student achievement. However, there was a positive trend (*r*'s = .05 to .40) in eight of the nine studies (Table 2.4).

Additional evidence supports this positive trend. In a study by Fortune (1967) of the behavior of student teachers presenting five to ten minute lessons to their classes, observers characterized the highest achieving teachers as both using more praise or repetition of a student's idea and integrating a student's idea into the lesson more frequently. However, the data were obtained from the descriptive reports of one observer, and he did not use a category system.

Using the data from Flanders' 6th grade study, Morrison (1966) compared the adjusted achievement scores of teachers who were in the upper third and bottom third in extended use of student ideas (3-3 cell). The results were significant at the .01 level on all seven subtests of the Metropolitan Achievement Tests used. However, Morrison used student as the sampling unit. Soar (1966) also compared the achievement scores of teachers who were extremely high and extremely low on this variable, obtained significant results in favor of the indirect teachers, and used students as the sampling unit. Although none of these three studies warrants inclusion in Table 2.4, they all support the positive trend for teacher use of student ideas.

Discussion. Although a great deal has been written about the importance of teacher use of student ideas (Cf. Flanders and Simon, 1969), the significance of this variable alone is not as strong as has been claimed. Judging by the available research, this variable is not as strong a predictor as "criticism or disapproval" but is more consistent a correlate than "praise."

## Table 2.4

## Acceptance of Student Ideas (Counting)

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Flanders, 1970<br>2nd, General<br>(15 teachers)<br>Two semesters | | Extended acceptance (3-3 cell), r = -.45 (sic) |
| Flanders, 1970<br>4th, Social Stud.<br>(16 teachers)<br>Two weeks | | Extended acceptance of student ideas (3-3 cell), r = .19 |
| Flanders, 1970<br>6th, General<br>(30 teachers)<br>Two semesters | | Extended acceptance of student ideas (3-3 cell), r = .30 |
| Flanders, 1970<br>7th, Social Stud.<br>(15 teachers)<br>Two weeks | | Extended acceptance of student ideas (3-3 cell), r = .40 |
| Flanders, 1970<br>8th, Math.<br>(16 teachers)<br>Two weeks | | Extended acceptance of student ideas (3-3 cell), r = .19 |

Table 2.4 (cont.) Acceptance of Student Ideas

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Perkins<br>5th, General<br>(27 teachers)<br>Two semesters | Factor III, Teacher leading recitation<br>teacher uses student idea | |
| | +[a] arithmetic reasoning | ns arith. fundamentals<br>ns reading<br>ns spelling<br>ns vocabulary |
| Soar, 1966 | Factor 8, Indirect teaching[b]<br>extended acceptance of student idea (.75)[c]<br>simple acceptance of student idea (column 3 of IA matrix) (.66) | |
| | | med. $r = .05$ |
| Wright and Nuthall, 1970<br>3rd, Science<br>(17 teachers)<br>Three ten-minute lessons | | Teacher repetition of student response, $r = .17$ |

---

[a] + refers to positive loading on a factor containing this variable; ns refers to no loading on a factor containing this variable.

[b] When the results of a factor analysis are presented in this and other tables, only those loadings relevant to the variable being considered are presented under the factor.

[c] This loading refers to the factor; it does not refer to any correlation with the student achievement measures.

However, the research in this area has only begun. Flanders has identified five different types of behaviors which might be classified as category 3. Frequencies for these smaller units (or subscripts of the major category) may yield higher correlations than category 3 taken as a whole. With Richard McAdams and Edward Crill of Temple University, this reviewer has recoded the audiotapes and transcripts made by Wright and Nuthall (1970) using the Expanded Interaction Analysis System developed by Amidon et al. (1969). In this system, category 3 is subscripted into three teacher behaviors: acknowledging the student's idea by a few words, such as "okay," or repeating what the student said; summarizing two or more ideas; and generalizing a student idea to a new situation. We found that repeating and summarizing behaviors each had a correlation of about .4 with student achievement, but that category 3 as a whole yielded a correlation of only .18 with student achievement. This single study, the first one reported in which subscripts were used, suggests that there may be more merit in subscripting behaviors in category 3 rather than in treating the variables within this category as a single type of behavior.

The results obtained by Soar (1965) also suggest that the concept of use of student ideas should be explored further. In this study, the frequencies in both column 3 and cell 3-3 had very low zero order correlations with the achievement measures, and these behaviors did not load on a significant factor. However, a different behavior, which was recorded using a modification of OScAR--teacher encouragement of pupil's interpretation and generalization--did have a positive, significant, zero order correlation with arithmetic achievement. Although these two types of behaviors both appear to involve teacher use of students' ideas, they were uncorrelated. Such results suggest that the concept of teacher use of student ideas is a complex one, deserving of more intensive future research.

## Combined or Unique Measures of Teacher Approval

Table 2.5 was created to include combined measures of teacher approval which do not fit easily into the above tables on non-verbal approval, praise, or use of student ideas. Teacher "indirectness" refers to the combined percentage of teacher behaviors in category 1 (acceptance of student feeling) plus

Table 2.5

Combined or Unique Measures of Teacher Approval (Counting)

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Flanders, 1970<br>2nd, General<br>(15 teachers)<br>Two semesters | | Indirectness (Columns 1 & 2 & 3), $r$ = -.04<br>General indirectness (Columns 1 & 2 & 3 & 4), $r$ = .05 |
| Flanders, 1970<br>4th, Social Stud.<br>(16 teachers)<br>Two weeks | | Indirectness (Columns 1 & 2 & 3), $r$ = .12<br>General Indirectness (1 & 2 & 3 & 4), $r$ = -.08 |
| Flanders, 1970<br>(6th, General<br>(30 teachers)<br>Two semesters | Indirectness (Columns 1 & 2 & 3), $r$ = .37* | General Indirectness (1 & 2 & 3 & 4), $r$ = .25 |
| Flanders, 1970<br>7th, Social Stud.<br>(15 teachers)<br>Two weeks | | Indirectness (Columns 1 & 2 & 3), $r$ = .41<br>General Indirectness (1 & 2 & 3 & 4), $r$ = .25 |
| Flanders, 1970<br>8th, Math.<br>(16 teachers)<br>Two weeks | | Indirectness (Columns 1 & 2 & 3), $r$ = .30<br>General Indirectness ( 1 & 2 & 3 & 4), $r$ = .45 |

Table 2.5 (cont.) Combined or Unique Measures of Teacher Approval

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Medley and Mitzel, 1959<br>3rd thru 6th, Reading<br>(49 teachers)<br>Two semesters | | emotional climate (refers to teacher and pupil supportive and reproving behavior), $r = .20$ |
| Penny, 1969<br>8th and 9th, Social Stud. and English<br>(32 teachers)<br>Two 45-minute sessions | Percent of times teacher followed a student response by the use of two or more reinforcing statements. (Such behavior would be classified as "extended indirect" in the Flanders Interaction Analysis matrix.)<br>Eight high-achieving and eight low-achieving teachers were compared on two independent occasions.[a]<br>$F = 7.0$* (August subsample)<br>$F = 4.2$* (total subsample) | $F = 1$ (June subsample) |
| Thompson and Bowers, 1968<br>4th, Vocab. and Social Stud.<br>(15 teachers)<br>Two semesters | | teacher supportiveness (similar to emotional climate studied by Medley and Mitzel)<br>dichotomized sample<br>$F<1$ (word meaning)<br>$F = 2.0$ (social studies)[b] |

---

[a] Results of videotape analysis of high-achieving and low-achieving teachers in each sample. Complete report does not give number of teachers studied, statistical procedures, or level of significance. A + was used to indicate that this behavior occurred more frequently in the high-achieving teachers, ns indicates that this behavior did not discriminate between the extreme samples.

[b] Mean scores not given in available report.

category 2 (praise and encouragement) plus category 3 (use of student ideas) in the Interaction Analysis coding system developed by Flanders (1965). Teacher "general indirectness" refers to the combined percentage of teacher behaviors in the above three categories plus those in category 4 (teacher questions).

Of the eight studies for which the statistical significance of the results can be assessed, only two yielded significant results (Flanders, 6th grade, 1970; Penny, 1969). The results for the sixth grade sample studied by Flanders add little to what is already known, because the significant correlation for "indirectness" ($\underline{r} = .37$) is almost identical to that obtained when praise alone was studied ($\underline{r} = .36$). Penny's finding (1969) that teacher use of multiple reinforcers discriminated between extremes of his sample is difficult to interpret because the complete report did not provide the operational definition of "reinforcer." A reinforcer might be restricted to praise, or it might include any or all of the five aspects of "use of student's ideas" (see above).

Although only one of the 11 correlations in this area was significant (Table 2.5), nine of them were positive. The positive correlations for "indirectness" ranged from .12 to .41; for "general indirectness," from .05 to .45. It was noted above that the range of positive correlations for praise was .14 to .49, and for use of student ideas, from .05 to .40. The similarity of these ranges suggests that little is gained by combining variables into measures of indirectness, general indirectness, "emotional climate" (Medley and Mitzel, 1959), or "supportiveness" (Thompson and Bowers, 1968).

Discussion. The overall conclusion is that combined measures of teacher approval such as indirectness yield weak but consistent correlations with student achievement. Gross measures of teacher supportiveness or indirectness are not as sensitive as measures of teacher affect which focus on contextual events, preceeding and subsequent events, and specific types of affect.

## Ratio of Teacher Approval to Teacher Disapproval Statements

In contrast to the few investigations in which praise or the use of student ideas was studied, 16 investigations were found in which a ratio of teacher approval to teacher disapproval either was found to be directly related to achievement or was used as part of a composite of teacher behaviors (Table 2.6). In twelve of these studies the i/d ratio (Flanders, 1965; Amidon and Flanders, 1967) was used. This ratio is formed by dividing the frequencies of teacher behaviors in categories 1 and 2 and 3 (see above) by the frequencies in category 6 (gives directions) plus category 7 (criticizes). The studies by Flanders (1965, 1970) conducted in the seventh and eighth grades are identical, except that the method of analysis differs. In the 1965 report, the 7th and 8th grade teachers were divided into two groups, and a critical ratio was computed using students as the sampling unit; in the 1970 report, linear correlation was used with class as the sampling unit.

One of the advantages of the Interaction Analysis system is that it yields a 10 X 10 matrix containing 30 cells taken to indicate warm, supportive, or "indirect" teacher behavior (columns 1, 2, and 3) and 20 cells taken to indicate critical, controlling, or "direct" teacher behavior (columns 6 and 7). Although the first i/d ratio (Flanders, 1965) was the ratio of frequencies in these two sets of cells, investigators have formed other ratios using selected cells within the total array of indirect and direct behaviors. In 14 of the 16 studies summarized in Table 2.6, at least one of three i/d ratios was used to describe teaching: the i/d, the i/d 8-9, and the extended i/d (Table 1.1 and Figure 1.1).

The use of different i/d ratios makes comparison between these studies difficult. Because there has been little research on the correlation of these i/d ratios, it is possible that if the investigators had used different i/d ratios, they might have obtained different results. In four of the five studies which used two i/d ratios to describe teaching, the results apparently would have been the same using either i/d ratio (Soar, 1966; Snider, 1966; Furst, 1967; Powell-fourth grade, 1968). In the fifth investigation, the study of third grade teachers by Powell (1968), different teachers would have been classified as direct or indirect if only the i/d ratio or the i/d 8-9 had been used in place of his composite score.

The IA system was not used in two of these studies. In one (Anthony, 1967) a ratio was formed of instances of positive affect to observed total affect. Such a ratio appears similar (if not identical) to the i/i+d ratio, which is frequently used

Table 2.6

Ratios of Teacher Approval Statements to Teacher Disapproval Statements (Counting) (i/d Ratios)

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Anthony, 1967<br>5th, General<br>(21 teachers) | Items in summed composite[a]<br>Ratio of positive affect to observed total affect, $r = .48$ for summed composite with achievement | |
| Birkin, 1967<br>5th, Reading<br>(34 teachers)<br>20 weeks | | i/d; i/d for row 8,<br>$r$ or F not given<br>Author states that trend was positive but ns |
| Cook, 1967<br>10th, Biology<br>(8 teachers)<br>Two semesters | | i/d for discussion, med. rho = .09<br>i/d for laboratory work, med. rho = .07 |
| Flanders, 1970<br>2nd, General<br>(15 teachers)<br>Two semesters | | i/d<br>$r = -.03$ |

---

[a] Of the 14 variables in the composite, only those relevant to positive affect or positive support are included here.

Table 2.6 (cont.) Ratio of Approval to Disapproval Statements

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Flanders, 1970<br>4th, Social Stud.<br>(16 teachers)<br>Two weeks | | i/d<br>$r$ = .33 |
| Flanders, 1970<br>6th, General<br>(30 teachers)<br>Two semesters | | i/d<br>$r$ = .12 |
| Flanders, 1965<br>7th, Soc. Stud.<br>(15 teachers)<br>Two weeks | i/d<br>teachers split into two groups according to i/d, CR = 5.02** | |
| Flanders, 1965<br>8th, Math.<br>(16 teachers)<br>Two weeks | i/d<br>teachers split into two groups according to i/d, CR = 3.42** | |
| Flanders, 1970<br>7th, Social Stud.<br>(15 teachers)<br>Two weeks | | i/d<br>$r$ = .47 ($p < .10$)<br>(NB: study identical to Flanders' 1965 study except that analysis was different) |
| Flanders, 1970<br>8th, Math.<br>(16 teachers)<br>Two weeks | | i/d<br>$r$ = .41<br>(NB: study identical to Flanders' 1965 study except that analysis differed) |

Table 2.6 (cont.) Ratio of Approval to Disapproval Statements

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Furst, 1967<br>10th and 12th, Social Studies<br>(15 teachers)<br>four 1-hr. lessons | teacher composite score on (a) extended i/d ratio, (b) i/d ratio for teacher responses to student talk, and (c) extended pupil talk<br>trichotomized sample, F = 3.90* | rho between (a) or (b) or (c) or composite score, not significant |
| Hunter, 1968<br>Emotionally handicapped children, ages 8 to 14, General<br>(11 teachers)<br>Two semesters | indirect/direct ratio obtained using modification of Withall system<br>med. r = .62* | |
| LaShier, 1967<br>8th, Biology<br>(10 teachers)<br>Six weeks | i/d ratio<br>tau = .59** | |
| Powell, 1968<br>3rd, Reading and Arith.<br>(9 teachers)<br>Two semesters | composite scores using seven variables indicating indirectness. These included i/d ratio, i/d ratio for teacher response to student talk, and extended i/d ratio.<br>Teachers divided into two samples for analysis<br>F = 10.68** (arith.) | F = 1.30 (reading) |

Table 2.6 (cont.) Ratio of Approval to Disapproval Statements

| Investigator | Significant Results | Non-Significant Results |
|---|---|---|
| Powell, 1969<br>4th, Reading and Arith.<br>(17 teachers)<br>Two semesters | same composite scores as Powell (above). Students same as above, but teachers were new. | $F < 1$ (reading)<br>$F < 1$ (arithmetic) |
| Snider, 1966<br>12th, Physics<br>(17 teachers)<br>Two semesters | | i/d ratio<br>analysis of extreme teachers<br>T-ratios on three criterion measures ns and quite small |
| Soar, 1966<br>3rd thru 6th, General<br>(55 teachers)<br>Two semesters | Factor 8, Indirect Teaching | i/d for responses to student talk (.49)<br>extended elaboration of student idea (.75)[a]<br>med. $r = .05$ |
| Torrence and Parent, 1966<br>(1st study)<br>7th thru 12th, SMSG Math<br>(10 teachers)<br>Two semesters | | i/d<br>rho = -.08 |

---

[a] Only factor loadings relevant to this variable are presented.

in place of the i/d ratio. The i/i+d ratio is formed by dividing the indirect teacher behaviors by the sum of all indirect and direct behaviors and is used to obtain a more normal distribution in cases in which there are few observed direct behaviors. In the other study (Hunter, 1968) the observational system devised by Withall (1949) was used, and the reviewer formed an indirect/direct ratio using the data available in Hunter's dissertation. This ratio appears identical to one which would be obtained using the IA system.

Results. It is difficult to present a simple overall summary of the results (Table 2.6) because of the variety of indirect to direct ratios used, and the variety of statistical procedures employed. One specific difficulty is that in four investigations both inferential and correlational procedures were used, and these procedures yielded different results (Flanders, 7th grade, 1965 and 1970; Flanders, 8th grade, 1965 and 1970; Furst, 1967; Soar, 1966).

Of the 13 studies which employed linear correlations in the study of an i/d ratio, significant results were obtained in three (Anthony, 1967; Hunter, 1968; LaShier, 1967). However, the results obtained by Anthony were part of a factor, and specific information on the i/d ratio cannot be obtained. The other two studies have questionable generalizability because Hunter studied educationally handicapped children, and LaShier studied student teachers instructing 8th grade students in a university laboratory school, using BSCS materials normally used in 10th grade classes. When the trend alone is considered, there were positive correlations in 11 of the 13 studies ($r$'s = .09 to .62) (Anthony, 1967; Birkin, 1967; Cook, 1967; Flanders, 4th, 6th, 7th, and 8th grade, 1970; Furst, 1967; Hunter, 1968; LaShier, 1967; Soar, 1966). Negative correlations were obtained in two studies, but these were rather small (Flanders, 2nd grade, 1970; $r$ = -.03; Torrance and Parent, 1966, rho = -.08).

Of the seven studies in which inferential statistics were employed to analyze extreme groups, a dichotomized sample, or a trichotomized sample (Table 1.6), significant results were obtained on at least one criterion measure in five studies (Flanders, 7th and 8th grades, 1965; Furst, 1965; Powell, 3rd grade, 1968; Soar, 1966). Non-significant and weak (e.g., $F < 1$) results were obtained in two studies (Powell, 4th grade, 1968; R. Snider, 1966). However, the significant effects in four of the five studies have questionable generalizability to the population of teachers because student was the sampling unit (Flanders, 7th and 8th grades, 1965; Powell, 3rd grade, 1968; Soar, 1966).

Discussion. The use of an i/d ratio to predict student achievement appears to yield consistent but weak results. The results are stronger when inferential statistics are used, but in these studies the data will have to be reanalyzed using class as the sampling unit before we can comment on the results. In addition, the results obtained when an i/d ratio is used do not differ appreciably from those obtained when other affective variables such as teacher criticism or teacher use of student ideas are taken singly. Of all the affective variables studied to date, criticism appears to yield the strongest results.

Even these sixteen studies do not reveal the whole picture on the predictive power of an i/d ratio. A variety of i/d ratios could have been computed in all these studies, and some form of i/d ratio might be consistently more predictive or differentiating than another. Indirect/direct ratios could have been computed in two additional studies (Perkins, 1965; Wallen, 1966), but the investigators did not do so, and the data for computing such ratios were not presented in the final report.

## Summary

In this chapter on variables related to teacher approval and disapproval, process-product relationships were reviewed in six categories of teacher behavior: criticism and control, non-verbal approval, praise, use of student ideas, indirectness, and indirect/direct ratios. In none of thee categories were there significant results formore than half the studies, but there were consistent positive trends for use of student ideas, indirectness, and indirect/direct ratios, and a consistent negative trend for criticism.

One point in the discussion of the results on each category was that specific types of praise, use of student ideas, criticism, or control yielded higher correlations than the entire category. Unfortunately, there were too few studies on these specific types to warrant conclusions. However, there may be value in expanding category systems to code specific forms of criticism or praise. Such expansion could focus on the intensity of the behavior, the context in which it occurred, and the events which preceded and followed the teacher behavior.

Although such expansion of category systems seems necessary for enhancing our understanding of those teacher behaviors which are related to student achievement, expanding the number of correlations which are computed also increases the probability of obtaining significant results by chance. This problem might be solved by greatly increasing the number of classrooms observed and using data reduction procedures, but the administrative problems and the expense of using observers currently preclude such arrangements. The best hope, at this time, may lie in increasing the number of investigations.

# Chapter Three

## Teacher Cognitive Behaviors

There has been much less systematic observation of the cognitive aspects of instruction than of affective aspects, and the observational measures developed by the different investigators are much more difficult to compare. Only 12 studies are reviewed in this chapter, compared to 30 studies in the chapter on affective behaviors. (However, there is a large body of research on cognitive teacher behaviors in which rating scales were used to estimate or evaluate specific cognitive behaviors. The most significant results of these studies are summarized in the Appendix.) Twelve studies represent only those investigations for which cognitive measures were developed and included in the analysis. Far more than 12 studies could be reported if the original investigators were to reanalyze their data. For example, in all studies in which IA was used, a cognitive measure could be developed by including the frequency or percentage of student predictable talk (Category 8) and student nonpredictable talk as part of the analysis. Unfortunately, few of the investigators who have used IA have attempted such analyses.

The variables in the cognitive area were grouped into six categories:

teacher questions-classified into two types.
teacher questions-classified into more than two types.
probing.
structuring.
task-oriented.
clarity.

The major emphasis is given to the first two categories because most of the studies fall in these two categories. The discussion on probing, structuring, task orientation, and clarity is primarily exploratory because very few studies are available in these areas. All of the above categorization is tentative; the reader is again encouraged to revise these categories as he reads this chapter or reads new studies.

Reviewers of the research on teaching behaviors (Medley and Mitzel, 1963; Amidon and Simon, 1965; Biddle, 1967; Meux, 1967; Lawrence, 1966; Campbell, 1968; Nuthall, 1970; Flanders and Simon, 1969; Rosenshine, 1970) have noted that most studies of teaching behaviors emphasized affective interactions; the cognitive aspects of teaching (e.g., the ability to explain new material, the effectiveness of various types of questions) received comparatively little attention. Gage observed that in research on teaching for cognitive objectives, "We have had relatively little of the . . . experimental or correlational work that can be found in relative abundance in research on the social and emotional phenomena found in classrooms" (Gage, 1966, pp. 32-33).

There are several possible reasons for this neglect. One, educational researchers have no analagous discipline to draw upon in developing observable cognitive variables. Research on child development, group dynamics, and experimental psychology can be used to discuss and code techniques of approval and disapproval, the cognitive interactions have not been developed in any discipline.

Second, although there has been a great deal of experimental research on cognitive variables in educational psychology, and such experiments appear throughout textbooks on educational psychology, few of these experimental variables appear in the classroom observational systems which have been developed. This neglect is probably not due to any preference of the researchers; rather, they may be unable to translate experimentally developed variables into a classroom grammar. For example, Ausubel (1963) has investigated the importance of the stability and clarity of cognitive structure by inserting "advance organizers" before a reading selection. Although Ausubel demonstrated the usefullness of a concept of cognitive structure, an investigator of classroom instruction cannot determine whether a teacher is adding organizers before the lesson--or during, or after the lesson--because the coding instructions needed to identify these behaviors have not been developed. In sum, until researchers can label the behaviors they observe, they cannot study either specific cognitive behaviors or the relationships between the behaviors and subsequent achievement.

Affective variables may also be easier to code because they are more independent of a person's previous cognitive experience. Statements like "Shut up and sit down" and "Excellent" are relatively clear, and we do not have to assess the nature of the audience before we code them. But the question, "How much is two and two?" is more difficult to classify. In a sixth grade classroom, we would feel confident classifying it as factual recall. But what if the question were asked of a five-year-old? This question might require convergent thinking or factual recall, depending upon the student's previous experience.

Developers of aptitude tests have avoided the problem of context by using puzzles and materials which are relatively unknown and content-free. But categories of questions developed in this context are not easily applied to the classroom where the questions, by design, are related to previous experience.

Because of the problem of context, it has been difficult to develop an observational system into which questions can be categorized reliably and meaningfully. Dichotomous classifications such as "narrow" and "broad," (Amidon and Flanders, 1967), "questions about content" and "questions that stimulate thinking," (Perkins, 1964), or "convergent" and "divergent" (Medley, et al., n.d.) appear to be oversimplifications of an area as complex as questioning, and they lead to different interpretations by different investigators. For example, Medley, et al., (n.d.) said that a divergent question admits of more than one answer, and therefore, "Name one of the four freedoms" is a divergent question. Other investigators would probably modify these instructions. Classification systems which divide questions into more than two types seem necessary.

Investigators whose systems for coding questions have been more elaborate have been forced to use transcripts or tape recordings of the class proceedings as the source for coding to allow coders the extra time necessary to categorize the behaviors (e.g., Bellack et al., 1966; Solomon et al., 1964). But even in these situations it has been difficult to develop categories whose boundary lines are clear.

One example of difficulty of doing research on the cognitive aspects of instruction is the large variance among the investigators in the systems they developed to quantify cognitive interchanges. Some categorized questions, some classified statements, and others quantified combinations of statements and questions. Consequently, comparison and synthesis of the results are particularly difficult.

Finally, the selection of an appropriate unit of measure (see Biddle, 1967) is even more difficult in studies on cognitive aspects of instruction. The simplest unit of measure or analysis is time, and Flanders' (1965) interaction analysis system with its three second rule has been used to code the cognitive level of classroom interaction in most of the studies reported in this chapter. Other investigators have attempted to develop cognitive units under which the frequency of events is recorded. These investigators have developed complex units such as a "move" (Bellack et al., 1966), a "venture" (Smith et al., 1967) or an "episode" (Conners and Eisenberg, 1966). Unfortunately it is much more difficult to train raters in the use of these complex units than it is to train them to use time as a unit, and frequently typescripts need to be transcribed before satisfactory inter-rater reliability can be obtained. Perhaps because of these difficulties, I found only three process-product studies which used a "move," a "venture" or an "episode" as the analytic unit.

## Types of Questions

The classification of questions and cognitive aspects of classroom interaction has been a difficult task. Investigators have differed widely in the types of questions chosen for analysis, whether questions were classified alone or as part of a larger unit, and the statistical treatment of the data. There is so much overlapping across investigations in procedures that it is particularly difficult to synthesize the results.

### Teacher Questions--Classified Into Two Types

Most of the investigators who studied teacher questions classified them into two types (Table 3.1). In general, the investigators distinguished between factual questions and those requiring thought, but the distinctions differed from study to study. It is impossible to determine with certainty whether the higher-level questions identified by Kleinman (1964), for example, differ from those identified by Spaulding (1965) or Wright and Nuthall (1970). Even when two or more investigators stated that they coded "divergent" questions, they may have used different operational definitions. Even if the definitions were explicitly and clearly given in the reports, we still would not know what modifications the observers made as they attempted to code the questions which teachers actually asked.

Harris and his associates (Harris and Serwer; 1966, Harris et al., 1968) used an observational system developed by Medley and Smith (1968). "Meaningful interchanges" are those which require a student to interpret a "word, sentence, or other symbol," "form interchanges" require only that a student recognize the symbol. These distinctions refer only to the teaching of reading. No further elaboration of these distinctions was given in the final report.

Kleinman (1964) classified questions into "low level" and "high level," with three subcategories within each type. Low level questions were classfied as "neutral," "rhetorical," or "factual." High level questions were further classified as "clarifying" (e.g., "What do you mean by friction?"), "associative," (e.g., "How do you compare the bird brain and the human brain?"), and "critical thinking," (e.g., "What are you basing your opinion on?"). Although Kleinman had the data to compare teachers on six types of questions, the only comparisons she made were between teachers who were extreme in low level or high level questions.

Perkins (1965) did not elaborate upon his definitions for two types of questions: questions about content, and questions to stimulate thinking (e.g., Why? How?).

Spaulding (1965) defined the eliciting of a specific answer as both containing recall questions and "giving mental arithmetic problems." "Open-ended questions" were those which elicited "judgment, opinion, interpretation, hypothesis, or prediction." It was impossible to determine from the definitions whether arithmetic word problems would be classified as "open ended" or "specific." Presumably they are "specific," because they contain an answer the teacher has in mind; however, the questions also may involve judgment and interpretation. Questions regarding children's interests are also open-ended, and in the examples, Spaulding used words such as "imagine," "what would the people feel?" and "can you tell us some interesting things."

Thompson and Bowers (1968) did not provide definitions or examples for convergent or divergent questions, but referred to an early form of OSCAR (Form 2V). Wright and Nuthall (1970) provided no definition or example of closed or open questions.

Soar (1966) used two categories ("teacher encourages factual answer," and "teacher encourages interpretation, generalization, solution") to categorize either the teacher's questions or the teacher's

Table 3.1

Types of Teacher Questions: Two Classifications

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Harris and Serwer, 1966<br>1st - Reading<br>(48 tchrs)<br>Two semesters | | r (observer counts)<br>percent form interchanges results not given, (presumably) ns<br>percent meaningful interchanges med r = -.17 |
| Harris et al., 1968<br>2nd - Reading<br>(38 tchrs)<br>Two semesters | | r (observer counts)<br>percent form interchanges results not given (presumably) ns<br>percent meaningful interchanges med r = -.11 |
| Kleinman, 1964<br>7th, 8th-Science<br>(6 of 23 tchrs)<br>Cross-sectional study | t-test (observer counting) high level vs low level questions<br>for students classified as high ability t = 5.02** | t-test (observer counting) high level vs low level questions<br>for students classified as average ability t = 1.29<br>for students classified as low ability t = 0.58 |
| Perkins, 1965<br>4th - General<br>(27 tchrs)<br>Two semesters | | r (observer counting)<br>teacher asks questions about content no loading on any factor containing student gain<br>teacher asks questions to stimulate thinking (e.g., why? how?) no loading on any factor containing student gain |

Table 3.1 (cont.)

Teacher Questions: Two Classifications

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Soar, 1966<br>3rd thru 6th -<br>General<br>(55 tchrs)<br>Two semesters | r (observer counting)<br>Factor 9: Unnamed[a]<br>teacher encourages answering factual questions (.30)[b]<br>teacher encourages answering question on interpretation, generalization, solution (.61)<br>r = 29* (arith concepts) | r (observer counting)<br>med r = .15<br>Factor 5: Unnamed<br>teacher encourages factual answer (-.50)<br>med r = -.03 |
| Spaulding, 1965<br>4th and 6th -<br>Reading and Mathematics<br>(21 tchrs)<br>Two semesters | r (observer counting)<br>Component 6: Businesslike[a]<br>teacher behavior<br>eliciting response in an open ended way (-.70)[b]<br>regarding child's interests, interpretations, or experiences (-.59)<br>r = .44* (reading)<br>r = .39[c] (mathematics) | r (observer counting)<br>Component 5: Calm acceptant teaching<br>eliciting answer teacher has in mind (.75)<br>regarding factually reported subject matter (.74)<br>regarding materials, resources, books, materials (-.53)<br>r = -.04 (reading)<br>r = .38[c] (mathematics) |

---

[a] Only variables relevant to this category are given as factor loadings.

[b] Factor loading, not correlation.

[c] $p < .10$

Table 3.1 (cont.)

Teacher Questions: Two Classifications

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Thompson and Bowers, 1968<br>4th-Vocab. and Social Studies<br>(15 tchrs)<br>Two semesters | F (observer counting) | F (observer counting) |
| | teacher questions classified on "convergent-divergent continuum" which was not further explained. This continuum probably refers to classification of questions as "convergent" or "divergent." | |
| | F = 4.56* (word meaning) | F = 1.90 (social studies) |
| Wright and Nuthall, 1970<br>3rd-Natural Science<br>(17 tchrs)<br>Three 10-minute lessons | | $r$ (typescript counting)<br>closed questions<br>$r = .31$<br>open questions<br>$r = -.08$ |

responses to student answers. Soar added (personal communication) that these categories refer to a teacher's questions and to his responses to students.

The major conclusion which I derive from Table 3.1 is that the simple categorization of questions into two types and the correlation of frequencies of these types with class residual mean achievement scores has not yielded significant or even consistent results. These non-significant results are puzzling. One would expect that the frequency of questions that encourage pupils "to seek explanations, to reason, to solve problems" (Perkins, 1965), or the frequency of questions related to interpretation (Harris and Serwer, 1966; Harris et al., 1968) would be consistently related to achievement.

These non-significant results have been experimentally replicated. Hutchinson (1963) ran an experiment in which four teachers taught the same material to two matched groups of seventh grade pupils. The instructional period was three weeks, or 15 fifty-minute lessons. After the first series of lessons, the teachers were given special training to increase their use of convergent, evaluative, and divergent questions (Gallagher and Aschner, 1963). They then taught the same material a second time to new groups of pupils. All class sessions were tape-recorded, and the frequency of use of different types of questions was tallied. These tallies indicated that the teachers used more high-level questions (i.e., convergent, divergent, and evaluative) when they taught the lessons a second time. Although the pupils who were taught the second series of lessons showed significantly more growth on some of the creativity tests, the two groups' mean scores on the achievement tests were almost identical.

Similar results were obtained by Miller (1966), although each question was not specifically categorized. Instead, all teacher statements were classified as "directive" or "responsive" according to an elaborate coding system. Under the responsive mode the teacher asks more high-level questions and elaborates pupil responses. In this experiment, each of four teachers taught 10 thirty-minute lessons to two groups of pupils; one set of lessons using the responsive mode, the second using the directive mode. Systematic observation of the teachers' behavior indicated significant differences between their behavior in the two settings, although the teachers were less consistent in following the responsive model. There were no significant treatment effects as measured by two criterion tests, one on mastery of facts and the other on "higher understanding."

In both studies, the levels of the pupils' responses were also coded; when the teachers asked higher-level questions, the pupils responded with higher-level answers. This additional evidence of differences in levels of thought in the two conditions in each study makes the non-significant results on the achievement measures in the correlational studies even more puzzling.

The results suggest two conclusions: (1) no clear linear relationship has been found between the frequency with which the teacher used certain types of questions and the achievement of pupils, and (2) the experimentally increased use of specified procedures or types of questions has not resulted in significantly increased achievement. As Conners and Eisenberg suggest from their study of the effect of teaching behavior upon IQ growth of preschool children, "It may be the total pattern of intellectual stimulation rather than any specific adherence to...different patterns of questions, that is required to induce growth" (1966, p. 10).

Additional Analyses. Significant results were obtained in four of the studies in Table 3.1, and although the analytic procedures and observational category systems used in these studies are too diverse to permit any synthesis, the results may provide some suggestions for future research.

In one study (Spaulding, 1965), the frequency of teacher's open ended questions regarding a student's interests, interpretations, or experiences was negatively related to achievement, but the category system which Spaulding used is so unique that these results cannot be compared with those obtained in the other studies. Another investigator (Soar, 1966) found that teacher encouragement of factual answers and teacher encouragement of interpretation and generalization loaded positively on a factor which was significantly related to arithmetic achievement. But this result is also difficult to interpret because both teacher questions and teacher responses were counted as "teacher encouragement." A third investigator (Kleinman, 1964) found that students classified as "high ability" learned more with teachers who asked more "high level" questions, and although this trend was maintained for students classified as average ability and low ability, the results were not significant. Finally, Thompson and Bowers (1968) computed a ratio of convergent and divergent questions and classfied teachers as high, moderate, or low according to this ratio. They found that teachers who were moderate (i.e., asked a relatively equal number of convergent and divergent questions) achieved significantly greater growth in word knowledge than teachers in the other two categories.

Although no discernible trend is apparent from these results, the studies provide several suggestions for future research. First, the classification of questions into two types and the correlation of the frequency of each type with the mean class residual gain score has not been a profitable pursuit to date. It is possible that better results could be obtained if investigators included means of subgroups of learners in their analysis. The study by Kleinman (1964) is an example of focusing upon subgroups of learners classified by their IQ scores. Second, the classification of questions alone may not be sufficient. In the study by Soar, the teacher's questions and the teacher's responses were coded together when counting the frequencies of "teacher encouragement of factual answers" or "teacher encouragement of interpretation and generalization." Third, the joint classification of teacher's questions and the topic of their questions may be useful. Such an approach was used by Spaulding, and his results suggest that the topic of the question is as important as its type. Finally, Thompson and Bowers' use of a convergent-divergent ratio provides a useful and potentially fruitful alternative to the simple correlation of frequency counts of question-type with achievement. The possibility of non-linear relationships which is suggested in the study by Thompson and Bowers is elaborated below.

As it stands, the simple categorization of questions into two types and followed by the correlation of frequencies in these types with class residual achievement scores has not yielded significant or even consistent results. However, the procedures used in four of the studies which obtained significant results, although too varied for synthesis, suggest a variety of research procedures which might be used in future studies.

## Types of Questions--Multiple Classification

Only three studies were found in which questions (or types of teacher-student interactions) were classified into more than two types. These studies differed widely in design and focus. In the study by Solomon et al. (1963), the analysis was done from tape recordings, and each independent clause of teacher statements, questions, and feedback, and of student statements and questions was categorized. All six categories used for classifying teacher (or student) questions are given in Table 3.2. In the study by Furst (1967), the analysis of teacher cognitive behavior was based upon the data provided by Bellack et al. (1966). In Bellack's study, each line of transcript was coded as to the logical-substantive process which was occurring. Separate results were not presented for teacher and student talk. When lengthly segments of teacher or student talk occurred, the entire segment was usually coded as to its dominant logical process.

In the study by Conners and Eisenberg (1966), the unit of measure was "groups of episodes." An episode was defined as a change of topic, change of teacher's attention from one student to another, or any new element of the teacher's behavior. Groups of episodes were classified according to the "implicit goal which these activities were judged to serve"; groups of episodes were defined as activities. The significant differences in the frequencies of different types of activities among the three groups of teachers are presented in Table 3.2. The two types of activities which yielded significant differences among high-, middle-, and low-achieving teachers were those which focused on "intellectual growth" and "property and materials." The activities were not elaborately defined in the complete report. Activities in "intellectual growth" were defined as those which focused on "language, concept, or symbolic training; factual knowledge about the world; development of sensory abilities, etc." The variable property and materials was defined as activities involving "consideration for the well-being, rights, and property of others."

Significant results were obtained in all three studies. Conners and Eisenberg (1966) found that the highest-achieving teachers had significantly more interactions which focused on intellectual growth, and the lowest achieving teachers had significantly more interactions which focused on property and materials. No significant differences were obtained on the remaining types of activities (Table 3.2). Furst (1967) found that the highest-achieving teachers had a higher ratio of analytic and evaluative interchanges divided by empirical interchanges. Solomon et al. (1963) found that two types of questions--interpretive and factual--loaded on a factor significantly related to gain in comprehension. Unfortunately, three studies are a small sample, and it is difficult to develop any summary statement which includes all three studies.

Perhaps the best conclusion which can be reached in this section is that the use of observational systems which include multiple classifications of cognitive interchanges has consistently yielded significant results. These consistent results stand in sharp contrast to the inconsistent and non-significant results obtained when only two types of questions were classified. However, only a few investigators have studied multiple classifications of cognitive interchanges, and very little is known about the effectiveness of various forms of questions.

Non-linear analyses of questioning. Perhaps the optimal relationship between types of questions and pupil achievement is not linear. Four investigators have studied this possibility: Soar, 1966; Thompson and Bowers, 1968; Furst, 1967; and Solomon et al., 1963.

Soar did not have a variable based on the frequency of questions per se; however, he used frequencies in columns 3, 4, 8, and 9 to develop ingenious measures of inquiry and drill. Inquiry was defined as the sum of the 3-3, 4-4, 8-8, and 9-9 cells; that is, extended teacher behaviors of elaborating pupils' answers, and extended questioning, as well as extended pupils' answers to teacher questions, or extended pupil-initiated responses. This pattern of extended time spent in questioning, elaborating, and answering was taken to represent inquiry. Drill was identified by the tallies in the 4-8 plus 8-4 boxes; that is, pupils' answers to narrow teacher questions plus teachers' questions following pupils' responses.

Soar developed three measures from these combinations: (a) the amount of inquiry, (b) the amount of drill, and (c) an inquiry/drill ratio computed by dividing the frequencies of inquiry behaviors by the frequencies of drill behaviors.

Two of these measures loaded on Factor 3, Discussion versus Rapid Interchange, a factor which had a positive correlation with all achievement measures and significant correlations with vocabulary, reading, and arithmetic concepts. Inquiry itself was not on this factor, but the inquiry/drill ratio had a positive loading, whereas drill had a negative loading. Soar interpreted this finding as suggesting that a classroom which is high on this factor is not especially high on inquiry, but is quite low on drill activities.

Thompson and Bowers (1968) classified questions into those for which more than one answer was possible (divergent), and those for which only one answer was possible (convergent). They also classified teachers as high, moderate, or low on a "convergent-divergent continuum" and found, using analysis of variance, that teachers classified as moderate had pupils whose achievement was highest in a test on word meaning.

Table 3.2

Multiple Classification of Questions or Interactions

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Conners and Eisenberg, 1966<br>Preschool<br>(38 tchrs)<br>6 weeks | F (observer counting) trichotomized sample<br>Focus of teacher-student interchange:<br>intellectual growth $F = 6.04^{*}$[H]<br>property and materials $F = 12.10^{**}$[L] | F (observer counting) trichotomized sample<br>Focus of teacher-student interchange:<br>self concept $F = 2.35$[M]<br>creativity $F < 1$<br>manners $F = 2.03$[L]<br>obedience $F = 1.67$[M]<br>rights of others $F < 1$<br>physical motor activities $F = 1.47$[H] |
| Furst, 1967<br>10th and 12th - Social Studies<br>(15 tchrs)<br>Four one-hour lessons | F (transcript counting) trichotomized sample<br>ratio of typescript lines (of teacher or student talk) devoted to analytic (i.e. defining) and evaluative substantive logical processes to the number of lines devoted to empirical (factual) logical processes.<br>$F = 16.92^{**}$[H] | rho (transcript counting)<br>rho = .38 |

[H] High-achieving teachers had highest frequencies.

[M] Middle-achieving teachers had highest frequencies.

[L] Low-achieving teachers had highest frequencies.

Table 3.2 (cont.)

Multiple Classification of Questions or Interactions

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Solomon et al.,<br>1963<br>College evening school<br>American History<br>(24 tchrs)<br>One semester | | Factor 1: Permissiveness[a] vs Control<br>hypothetical ques. (-.78)[b]<br>opinion questions (-.70)<br>organizing ques. (-.66)<br>interpretation ques. (-.49)<br>non-specific ques. (-.67)<br>*r* = .19 (factual gain)<br>*r* = .32 (compre. gain) |
| | Factor 2: Energy versus Lethargy<br>interpretation ques. (.63)<br>factual questions (.49)<br>*r* = .44* (compre. gain) | *r* = .23 (factual gain) |

---

[a] Only factor loadings relevant to this variable are presented.

[b] Factor loading, not correlation.

One of the items in the cognitive composite developed by Furst (1967) was the ratio of analytic and evaluative to cognitive processes. She developed this ratio by using the descriptive data provided in the report by Bellack et al. (1966); the report gave the number of lines in the transcripts of the classroom interactions which Bellack and his associates coded as involving analytic, evaluative, or cognitive processes. Analytic refers to defining or interpreting the meaning of an item or statement; empirical includes fact stating or explaining the relationship between events; and evaluative deals with personal judgments and/or the reasons for the judgments.

Furst hypothesized that the superior teachers would show greater variety in their use of these processes; she computed the ratio of the two least frequently used to the most frequently used cognitive processes (i.e., the ratio of the lines devoted to analytic plus evaluative processes to the lines devoted to empirical processes). Inspection of the original data (Furst, 1967, p. 203) indicated that the three most effective teachers were significantly superior to the remaining teachers on the variable variety of cognitive processes.

Solomon et al. (1963) found that six of the seven types of questions loaded on Factor 1, labeled "permissiveness versus control." (Rated items were also included in developing the factors, which accounts for the discrepancy in the label). Although there was no significant linear relationship between teacher loadings on this factor and either of the achievement measures, teachers who were moderate on this factor had classes with significantly higher difference scores on the comprehension test.

In each of these four studies, the methods for analysis are quite different, although each method appears to have achieved limited success and appears useful for future research. Perhaps one of the more fascinating discoveries in reviewing these studies is the variety of procedures which different investigators have used. They have varied in their classification schemes, units of analysis, and statistical procedures. There is no simple way of testing which of these numerous combinations of procedures will obtain optimal results. It is possible that one set of procedures will be more effective in accounting for student achievement in one situation, and another set in another situation. But any set of "optimal" procedures will have to be replicated using another sample from the same population, and at this stage in our research such replication occurs infrequently.

## Other Cognitive Variables

The results on cognitive variables other than teacher questions are discussed below. These are: probing, structuring, task-orientation, and clarity. Too few studies have been conducted in these areas to permit any conclusions or synthesis of the results. However, all four variables appear worthy of future study, and the specific procedures used in each of the studies might be useful for future investigators.

### Probing

The results of three studies (see Table 3.3) suggest that there may be merit in investigating the teacher's cognitive response to student answers (Soar, 1966; Spaulding, 1965; Wright and Nuthall, 1970).

In the modified version of OSCAR 2V used by Soar (1966), teacher questions and statements were coded into three categories: (a) teacher encourages further answers to fact questions, (b) teacher encourages further explanations, and (c) teacher encourages inter-relationships, generalizations, and problem solutions. Only one of these three variables loaded on a significant factor. Teacher encouragement of inter-relationships and generalizations loaded on the unnamed Factor 9, which had a significant, positive correlation with achievement in arithmetic concepts, and positive, although not significant, correlations with the remaining product measures.

Teacher repetition, clarification, or use of pupil ideas may be another form of cognitive response to student answers. The only study in which teacher "use of student ideas" and teacher response by clarification were both included in the observational system (Soar, 1966) indicated a negligible correlation between these two behaviors. Perhaps these procedures are uncorrelated but equally effective means of achieving the same ends. In addition, Soar found that teacher use of inquiry, or the inquiry/drill ratio, was not related to the frequency of either type of cognitive response.

In the study by Spaulding (1965), a variable which has already been discussed under affective behaviors, "teacher elicits clarification in a non-threatening way," loaded on a component which was significantly related to reading gain, and nearly significant in mathematics gain. This behavior appears to be only

Table 3.3

Probing

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Soar, 1966<br>3rd thru 6th - General<br>(55 tchrs)<br>Two semesters | *r* (observer counting)<br>Factor 9: Unnamed[a]<br>Teacher encouragement of interpretation, generalization, and solution<br>*r* = .29*(arith.concepts)[b] | *r* (observer counting)<br>med *r* = .19 |
| Spaulding, 1965<br>4th and 6th - Reading and Mathematics<br>(21 tchrs)<br>Two semesters | *r* (observer counting)<br>Component 6: Businesslike, orderly teacher behavior[a] eliciting clarification in a non-threatening way (.36)[b]<br>regarding lack of knowledge (to boys) (-.65)<br>regarding lack of knowledge (to girls) (-.30)<br>regarding lack of knowledge (to class) (-.12)<br>regarding lack of attention ( .30)<br>*r* = .44* (reading | *r* (observer counting)<br>*r* = .39 (math) |
| Wright and Nuthall, 1970<br>3rd<br>Science<br>(17 tchrs)<br>Three ten-minute lessons | *r* (typescript counts)<br>redirects question *r* = .54*<br>teacher information following question *r* = -.52* | *r* (typescript counts)<br>alternative subsequent question *r* = -.40<br>reciprocates to extend *r* = .20<br>reciprocates to lift *r* = -.20 |

[a] Only loadings relevant to this category are presented.

[b] Factor loadings; not correlations.

tangentially related to probing because it is part of the molar behavior "disapproval." No behavior which might be labeled as probing appeared under the molar behavior "instruction."

In the study by Wright and Nuthall (1970), there were five categories used to code teacher cognitive responses to a student's answer. These included asking a question on the same cognitive level to the same student (alternative subsequent question) or to another student (redirects question), or asking questions at a higher thought level (reciprocates to lift). As noted in Table 3.3, the frequency of redirection of questions was significantly related to student residual mean achievement ($\underline{r} = .54$), but the other forms of cognitive response yielded mixed results.

The teacher's cognitive response to student answers is a particularly difficult area to investigate as shown in the varied correlations obtained by Wright and Nuthall (1970). The fact that significant results were obtained in all three studies on variables which have some approximation to "probing" suggests that there is merit in continuing to study the teacher's cognitive responses to student statements. But the category systems used in these three studies are so varied that no conclusions can be drawn with confidence. It is particularly unfortunate that there are so few studies in this area in view of the emphasis which "learning by inquiry" has received in curriculum courses.

## Structuring

The term "structuring" is used here to refer to four overlapping variables which were studied in at least one of the studies summarized in Table 3.4: The teacher comments made at the beginning or at the end of a lesson, and the teacher comments made before or after he asks a question. Although the effects of introductory and concluding statements have been investigated in laboratory studies using meaningful verbal material by Ausubel, Rothkopf (1970), R. Anderson (1970) and their associates, there has been little classroom research in this area.

Five low-inference studies were found in which variables similar to structuring were studied (Table 3.4). Crossan and Olson

## Table 3.4

### Structuring

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Crossan and Olson, 1969<br>6th and 12th - Special tests in verbal learning and symbolic learning<br>(6 and 35 tchrs)<br>Two ten-minute lessons | audiotape counting (significance tests not made)<br>clear signal when one part of lesson ended and another began<br>emphasis upon words to be learned | |
| Furst, 1967<br>10th and 12th - Social Studies<br>(15 tchrs)<br>Four one-hour lessons | F (typescript scoring)<br>moderate number of "teacher structuring lines" (coded by Bellack et al., 1966)<br>trichotomized sample (part of significant composite;"structuring" in itself not analyzed) | rho of achievement and deviation of structuring lines from mean = .48 |
| Penny, 1969<br>8th and 9th - Social Studies and English<br>(32 tchrs)<br>Two 45-minute sessions | F (typescript scoring)<br>verbal markers of importance (e.g., "now get this.")<br>F = 7.9* (1st sample)<br>F = 4.3* (2nd sample)<br>F =10.6** (total) | |

Table 3.4 (cont.)

Structuring

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Soar, 1966<br>3rd thru 6th - General<br>(55 tchrs)<br>Two semesters | *r* (observer counting)<br>Factor 3: Extended Discourse<br>extended lecture (.80)[a]<br>med *r* = .37* | |
| Wright and Nuthall, 1970<br>3rd - Science<br>(17 tchrs)<br>Three ten-minute lessons | *r* (typescript counting) | *r* (typescript counting) |
| | review at end of lesson<br>*r* = .67** | terminal structuring (structuring at the end of an episode)<br>*r* = .41 |
| | Teacher information following question<br>*r* = -.52* | structuring prior to a question<br>*r* = -.13 |
| | | review at start of second lesson<br>*r* = .18 |
| | | review at start of third lesson<br>*r* = -.08 |

---

[a] Only factor variables relevant to this table are presented.

(1969) recorded whether teachers gave a clear signal when one part of the lesson ended and another began. No one else studied explicit marking of transitions. The use of emphasis, which might be taken as a form of structuring, was investigated in two studies. Crossan and Olson counted teacher emphasis upon words to be learned, and Penny (1969) developed a category which he named "verbal markers of importance" (e.g., "Now this is important!").

Using the coded data provided by Bellack et al. (1966), Furst (1967) included the number of structuring statements which a teacher made as part of her "cognitive composite." In the coding system developed by Bellack et al., structuring referred to the initial statements of the teacher which serve to initiate or focus a teaching cycle (or move). These initial statements frequently precede a question. (Bellack did not make a separate count of structuring at the end of an interchange; such a modification was used by Wright and Nuthall in their investigation.) Furst used a unique procedure for determining "moderation" in structuring by assigning the lowest weight in a Fisher standard measure to the teacher who was closest to the mean of the 15 teachers in structuring. Teachers who deviated from the mean (regardless of the direction) were assigned higher weights according to the amount of deviation from the grand mean which they exhibited. The finding by Furst that the highest-achieving teachers were moderate in their use of structuring statements suggests that providing a moderate amount of structure was the most effective teaching procedure for those high school classes.

Soar (1966) believed that the positive relationship between steady-state lecturing (cell 5-5) and achievement reflects cognitive structuring activities on the part of the high-achieving teachers. Such a hypothesis cannot be investigated by an inspection of an IA matrix because both extended and relatively short lecturing would fit into the 5-5 cell. But Soar studied the original observer tally sheets and determined that four of the five highest teachers on his Factor 3 followed a pattern in which they lectured at most for 15 or 20 seconds, and then asked a question, and the pupils responsed. Such a pattern of posing a situation or providing limited units of information, asking a question, and responding to the question appeared to Soar to parallel the rationale of programed instruction. It is possible that both Soar and Furst have identified such a pattern in successful teachers. Gage and Unruh (1967) have also suggested a parallel between the structuring, soliciting, responding, and reacting pattern described by Bellack et al. (1966) and the sequence of frame-presentation, question, response, and reinforcement which appears in programed instruction.

The coding system used by Wright and Nuthall (1970) contained both original categories and a modification of the system developed by Bellack et al. (1966). The "structuring" statements used by Bellack were divided into "structuring prior to a question" and "terminal structuring,"or structuring at the end of a move. In addition, these investigators also studied review at the start of the second or third 10-minute lesson, and review at the end of a lesson. Another variable, "teacher information following question," refers to the teacher's use of additional statements after he has asked a question and before any student has answered. The significant negative correlation obtained for this variable (_r_ = -.52) might also be taken as a measure of the lack of clarity in the question. That is, structuring statements following a question may have been necessary because the students were unable to answer the question.

Of the five low-inference studies of variables which might be labeled structuring, all studies yielded positive results. These results were significant in three studies (Furst, 1967; Penny, 1969; Soar, 1966) and both significant and non-significant in one study (Wright and Nuthall, 1970); the level of significance was untested in one study (Crossan and Olson, 1969). In the study with both significant and non-significant results (Wright and Nuthall, 1970), review at the end of a lesson or an episode appeared to be more effective (_r_s = .67 and .41, respectively) than review at the start of a lesson or structuring prior to a question (_r_s = .18 and -.13). The studies by Furst (1967), Soar (1966), and Wright and Nuthall (1970) suggest that the amount of structuring before asking a question is an important variable, but it is difficult to determine from these studies what the optimum level is.

_Future Research_. The fact that significant results were obtained in all four studies for which statistical tests were run (see Table 3.4) clearly favors structuring as a promising area for future research. However, because the investigators studied different variables and used different statistical treatments of the data, any synthesis of the results appears premature. In addition, the definitions of structuring, to say nothing of advance organizers, are far from precise. Future studies in this area might focus on the effects of structuring at different places in the lesson, and the effects of structuring before or after a series of events (such as structuring following a question, or structuring at the end of a move or at the end of a lesson). One might also consider the relationship between structuring comments at the start of an episode (or move) and the quality of questions. Short structuring sentences before a question may facilitate achievement (Soar, 1966), but such statements may not be necessary if the questions are clear (Wright and Nuthall, 1970). The interaction of structuring statements and the clarity of questions might therefore be considered in future research. One way of measuring the clarity of questions might be to determine whether students answered a question the first time it was asked (Wright and Nuthall, 1970). Finally, in future research, it might be appropriate to consider non-linear analyses in addition to the more frequently used linear analysis.

## Task Oriented Behavior

The variable "task oriented, achievement oriented, or businesslike behavior" has primarily been studied using high-inference rating scales. The results of such studies are summarized in the appendix. However, two studies were found in which the counted behaviors of the teachers also appeared to suggest achievement- or task-oriented teacher behavior (Table 3.5). In the study by Conners and Eisenberg (1966), the most effective teachers had significantly more teacher-student interchanges which focused on intellectual content, and significantly fewer interchanges which focused on property and materials. Spaulding (1966) identified one of his components as "businesslike, insisting upon attention." The specific behaviors which comprise this component (and their loadings) are presented in Table 3.5. The businesslike teacher identified by Spaulding appears to be characterized by avoiding open-ended questions and instruction regarding student's interests, and by avoiding approval regarding personal interests or disapproval regarding lack of knowledge. Positive cognitive behaviors did not appear clearly on this factor. Spaulding (personal communication) chose the title of "Businesslike" for this component because of all the molar behaviors in his category system (that is, major categories into which all behavior was first coded, such as "approval," "disapproval," "instruction," and "listening,"), "instruction" had the highest loading on this component (.29). Within the subcategories of technique of instruction, the highest positive loadings were for "stating facts authoritatively, lecturing" (.25), and "eliciting the idea or answer that the teacher had in mind" (.20). However, such loadings are below the usual cutoff of .40 for selecting items which define a factor.

## Table 3.5

### Task Oriented, Achievement Oriented, Businesslike

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Conners and Eisenberg, 1966<br>Preschool<br>(38 tchrs)<br>6 weeks | F (trichotomized sample)<br>Number of interchanges which focused on:<br>intellectual growth $F = 6.04$*[a]<br>Propery and materials $F = 12.10$**[b] | F (trichotomized sample)<br>Number of interchanges which focused on:<br>creativity $F < 1$<br>manners $F = 2.03$<br>rights of others $F < 1$ |
| Spaulding, 1966<br>4th and 6th - Reading and Math<br>(21 tchrs)<br>Two semesters | Component 6: Businesslike, Insisting Upon Attention<br>approval in normal tone of voice[c] (.36)<br>approval using warm voice (-.39)<br>approval regarding personal interests of student (-.39) | |

---

[a] Highest-achieving group had highest frequency on this variable.

[b] Lowest-achieving group had highest frequency on this variable.

[c] All variables which loaded on this factor are included. Factor loadings are in parenthesis.

Table 3.5 (cont.)

Task Oriented, Achievement Oriented, Businesslike

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Spaulding, 1966 (cont.) | disapproval by commanding conformance (.41) | |
| | disapproval regarding lack of knowledge (-.65) | |
| | instruction to boys (-.42) | |
| | instruction in normal voice (.41) | |
| | instruction in warm voice (-.56) | |
| | eliciting verbal response in open-ended way (-.70) | |
| | instruction regarding student's interests (-.59) | |
| | $\underline{r}$ = .44* (math | $\underline{r}$ = .39$^{d}$ (reading) |

---

$^{d}p$ .10

In summary, both studies in Table 3.5 provide some support for the importance of achievement-oriented teaching. This variable received greater support from studies in which teacher behavior was rated, and these reusults are summarized in the Appendix.

## Clarity

The clarity of a teacher's presentation has been studied mainly by correlating observer ratings on this variable with measures of residual achievement (see Appendix). The results in such high-inference studies are most encouraging--significant results were obtained in eight of eight studies. But although there is strong support for the validity of clarity as a high-inference variable, its low-inference version is difficult to evaluate because only one study (Solomon et al., 1963) used a low-inference measure of this variable. In that study, the factor Clarity contained relatively low and negative loadings for "proportion of student interpretation of total student speech" (-.47) and "proportion of teacher interpretation of total teacher speech" (-.43). Interpretation is defined in a single sentence in the complete report: "Focus upon explicit attempt to understand, explain, (sic) course materials" (pp. 137-138). Apparently teachers high in clarity spent less time interpreting course materials. It is possible that such teachers were able to make a clear presentation the first time.

Perhaps additional low-inference components of clarity or intellecutal effectiveness are contained in the study by Wright and Nuthall (1970). They found that teacher "utterances" containing one question were positively and significantly related to achievement ($r$ = .54), whereas utterances with two or more questions or with teacher information following a question were each negatively related to achievement ($r$s = -.43 and -.52). An "utterance" was defined as a single teacher-pupil interaction. Thus, teachers who more frequently asked questions that were answered the first time were more effective, whereas those who more frequently had to ask a second question before receiving an answer, or who more frequently followed a question with a statement of their own, were less effective.

Summary

In contrast to the 30 studies on affective teacher behaviors which are summarized in Chapter 2, only 12 studies were located in which classroom observational category systems were used to code teacher behaviors which might be regarded as cognitive. In this limited group of studies, the greatest emphasis was on coding teacher questions, and other cognitive aspects of classroom interaction received relatively little attention.

In the area which was studied most frequently--the classification of questions into two types--there was no consistent linear trend favoring frequent use of questions classified as representing "higher" or "lower" cognitive processes. Significant results which were obtained on non-linear analyses, multiple classification of questions, probing, structuring, task orientation, and clarity can be considered only as suggestive for future research because too few investigators have focused on such variables, and their methods of research are too diverse to permit any synthesis of their results.

It is unfortunate that although cognitive achievement is one of the accepted goals of schooling, there have been so few studies which employed cognitive variables in their observational systems.

# Chapter IV

## Flexibility and Variety

Even though "flexibility" has long been honored as a characteristic of effective teachers (see Hammacheck, 1969), relatively few studies have employed this variable. A major difficulty is defining the variable. Two approaches have been used in counting teacher flexibility. One is to count variation in behavior without focusing upon specific behaviors; the other is to count frequency of variation in specific activities.

General Variation in Behavior In the studies on Table 4.1 general variation in teacher behavior was counted. Flanders defined flexibility as, "the arithmetic difference between the largest i/d ratio over all time use categories (e.g., discussion, administrative routine, new material, etc.) and the smallest i/d ratio for all time use categories" (Flanders, unpublished draft document). Soar (1966) defined flexibility as the number of cells in an IA matrix necessary to account for 60 per cent of the tallies. A teacher who used a large number of different cells in the 100-cell matrix would have a high flexibility score. Snider (1966) used the standard deviation of the i/d ratio in different activities such as lecture and discussion.

Of the eight studies in which general variation in behavior (or flexibility) was counted, none yielded significant results. Positive correlations were obtained in five of the seven studies for which the direction of the relationship could be determined (Table 4.1).

Variation in Specific Behaviors Other investigators studied variation or flexibility by focusing upon specific teacher behaviors. Three such studies are summarized in Table 4.2.

Anthony (1967) identified a number of measures of variation in specific activities through interviews with the teachers and observation in the classroom. The frequency counts were then converted to seven-point scales for use in the statistical analysis. Those variables which appeared on her single factor are presented in Table 4.2. It should be noted that variables reported in Chapter One from the same study by Anthony also loaded on this factor. Those variables included high positive affect and low negative affect analysis. Fourteen

Table 4.1

Flexibility

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Flanders, 1970<br>2nd - General<br>(15 tchrs)<br>Two semesters | | r (observer counting) flexibility--defined as largest i/d during instructional unit minus smallest i/d ratio.<br>$r = -.07$ |
| Flanders, 1970<br>4th - Social Studies<br>(16 tchrs)<br>Two weeks | | flexibility<br>$r = .46^a$ |
| Flanders, 1970<br>6th - General<br>(30 tchrs)<br>Two semesters | | flexibility<br>$r = .19$ |
| Flanders, 1970<br>7th - Social Studies<br>(15 tchrs)<br>Two weeks | | flexibility<br>$r = .37$ |
| Flanders, 1970<br>8th - Math<br>(16 tchrs)<br>Two weeks | | flexibility<br>$r = .43^a$ |

$^a p < .10$

Table 4.1 (continued)

Flexibility

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Snider, R., 1966<br>12th - Physics<br>(17 tchrs)<br>Two semesters | | U-test (observer)<br>a) range of four i/d measures<br>b) s.d. of i/d for different activities (e.g., lecture, discussion)<br>No U-tests significant on 3 criterion measures |
| Soar, 1966<br>3rd thru 6th - General<br>(55 tchrs)<br>Two semesters | | r (observer-IA)<br>Factor 5: Unnamed[a] flexibility (-.62)[b]<br>med r = -.02 |
| Vorreyer, 1965<br>5th - General<br>(14 tchrs)<br>Two semesters | | rho (counting)<br>flexibility defined as number of times teacher changes behavior<br>med rho = .20<br>variety defined as number of different behaviors in time interval<br>med rho = .25 |

---

[a] Only loadings relevant to "flexibility" are given in this table.

[b] Factor loading; not correlation.

variables appeared most promising, and these loaded on a single factor. These fourteen variables might be subdivided into three types: the variety of instructional materials and techniques, the frequency and variety of reinforcements used by the teacher, and the types of feedback available to teachers and students through testing. Variety is common to all three subdivisions. The variables referring to reinforcement have already been covered in Table 2.3 under "teacher praise." The behaviors relevant to variety of instructional materials and techniques are listed in Table 4.2.

The work of Furst (1967), Thompson and Bowers (1968), and Soar (1966) has previously been described in Chapter 3 under different sections. The studies and results are again described in this chapter because their results also appear to be relevant to consideration of variation in specific activities. The fact that the same set of variables can appear in more than one chapter is another indication of the lack of conceptual clarity in this relatively new undertaking.

In the study by Furst (1967), both student and teacher talk relevant to the subject area (compared to managerial talk) were classified into one of three major cognitive processes: analytic (or defining), evaluative, and empirical (fact stating and explaining). The highest frequency of cognitive interaction across the sample was on the empirical level. Furst reasoned that teachers who used a variety of cognitive processes might obtain greater achievement, and therefore she calculated (for each teacher) a ratio of the most frequently used process ( empirical) to the least frequently used processes (analytic and evaluative). These ratios were converted into standardized scores in which teachers who showed the greatest variation (had the lowest ratio) received the lowest scores.

Thompson and Bowers (1967) apparently computed a ratio of convergent and divergent questions for each teacher and classified the teachers as highly convergent, highly divergent, and moderate. Those classified as moderate apparently had the largest mixture of the two types of questions.

The results for all three studies are presented in Table 4.2; significant results were obtained in all cases. For Anthony, the factor on which these behaviors loaded was significantly related to student residual gain measures, although the individual weights for each of the variables was not given. In the study by Furst, the three teachers who obtained the highest residual gain scores also had significantly greater cognitive variation than the remaining teachers. In the study by Thompson and Bowers, the teachers who were moderate in divergent-convergent questions obtained the highest achievement in vocabulary (although differences were non-significant in social studies gain). However, none of the three investigators reported significant

Table 4.2

Variety and Variation

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Anthony, 1967<br>5th - General<br>(21 tchrs)<br>One semester | $r$ (observer and interview)<br>Factor 1 (14 variables)<br>(loadings not given)<br><br>variety in test form<br>distinct variety in objects handled by pupils<br>variety in objects handled by pupils<br>variety in observed teaching devices<br>number of three-dimensional displays<br>number of observed displays on academic subjects<br>novel or real-life displays in classroom<br>$r = .48+$ | |
| Furst, 1967<br>10th and 12th grades - Social Studies<br>(15 tchrs)<br>four one-hour lessons | F (typescript scoring)<br>trichotomized sample | F (typescript scoring) |
| | ratio of typescript lines of teacher or student talk using analytic (defining) and evaluative substantive logical processes to lines using empirical processes. | |
| | $F = 16.92{+}{+}^{H}$ | rho = .38 |

$^{H}$Highest achieving teachers had highest ratio.

Table 4.2 (continued)

Variety and Variation

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Thompson and Bowers, 1968<br>4th - Vocab. and Social Studies<br>(15 tchrs)<br>Two semesters | F (counting)<br>(trichotomized sample)<br>teachers classified as high, medium, or low according to questions on a "convergence-divergence continuum." | |
| | $F = 4.56^{+}$ (medium group highest)[a]<br>(Vocabulary) | F= 1.9<br>(Social Studies) |

---

[a] Teachers who were medium on this variable had classes with significantly higher achievement scores.

correlational results for their measure of variation. In the study by Anthony, teacher use of a variety of materials was only part of a factor which included other teacher behaviors such as affective responses to students; in the study by Furst, the rank order correlation was not significant; and no estimate of the correlation could be obtained from the study by Thompson and Bowers. Despite these limitations, the results are distinctive for their consistency.

The importance of variation in specific activities also receives some support from the study by Soar (1966) which is also described more fully in Chapter Three. Soar found that the most effective teachers had a higher ratio of "inquiry" to "drill" activities, although frequency of inquiry procedures itself was not a significant correlate, and drill was negatively related to achievement. ("Inquiry" and "drill" were defined by combining cells in the IA matrix in a procedure only used by Soar, to date). Soar's finding may suggest that the most effective teachers were moderate in their use of inquiry and low in their use of drill. Whether such findings can be taken as support for "variation in specific activities" is speculative, but these results are added in the hope that they may be of interest to investigators who are considering future correlational and experimental studies in this potentially fruitful area.

However, it is difficult to compare or attempt to synthesize these four studies because they again involve widely disparate observational systems and statistical procedures. Yet, all four studies suggest that variation in classroom activities or in cognitive processes may be an important correlate of student achievement.

The importance of variation in activities is also supported by two studies in which student attention was the criterion measure. Kleinman (1964) summarized a mimeograph report by Wilk et al. (1960), in which a modification of OSCAR was used as the observation instrument. According to Kleinman, Wilk found that "the amount of disruptive behavior in the classroom was negatively correlated with the amount and variety of classroom activities" (Kleinman, 1964, p. 37).

Using an observational category system, Kounin (1970) reported that the correlation between variety of school-unique activities (e.g., reading, arithmetic) and work involvement was .83 and .52 for two samples of 11 and 49 classrooms, respectively. The students were in the first and second grades. For nine classrooms of grades three through five, the correlation between seatwork variety and work involvement was -.67 (sic). Although Wilk et al. and Kounin used student attention to task as their criterion, other investigators found that observer ratings of student attention were significant and consistent correlates (rs = .39 to .62) of student achievement (Belgard et al., 1968; Hunter, 1968; Lahaderne, 1968; Morsh, 1956; Shannon, 1942).

## Summary

The studies on teacher flexibility have yielded two sets of results. When flexibility was defined as changes in all types of teacher behavior, none of the studies yielded significant results (Table 4.1). When flexibility was defined as variation in the teacher's cognitive behavior or the richness and variety of classroom materials and activities, then the results were consistently significant (Table 4.2).

# Chapter Five

## Amount of Teacher-Student Interaction

The emphasis in this chapter is on the amount, rather than the type of teacher-student interaction. The results of 13 studies are presented as teacher talk, student talk, and teacher-student interactions. This is a short chapter. The variables were not included as part of the chapter on cognitive variables (Chapter Three) because in the earlier chapter the focus was on the type of cognitive interaction, whereas in this chapter the focus is primarily upon the amount of interaction.

### Teacher Talk

Teacher talk (Table 5.1) has been studied by summing the frequencies of all behaviors labeled as teacher talk. The Flanders Interaction Analysis system has been used in the studies by Flanders (1970) and Sharp (1966). Teacher talk was determined by counting lines in transcripts in the study by Solomon et al. (1963) and in the study by Wright and Nuthall (1970). Wright and Nuthall defined teacher utterances as teacher statements or series of questions which are unbroken by student talk. Altogether nine studies are reported in Table 5.1.

In the eight studies for which correlations were available, the frequency of teacher talk yielded consistent positive correlations which were low and not significant. A ninth study might be added to this group by including the study by Soar (1966). In this study, a variable named "steady state lecture" referred to three seconds of teacher talk followed by an additional three seconds of teacher talk (Cell 5-5). "Steady state lecture," however, is only a part of the total teacher talk. This variable had a loading of .80 on Factor 3, a factor which had a median correlation of .27 ($p < .05$) with the five measures of student achievement, and was significantly correlated with residual student achievement in vocabulary, reading, and arithmetic concepts.

The single negative relationship between teacher talk and student achievement was obtained in the study by Perkins (1965) (see Table 5.1), in which total teacher talk (or "teacher lectures") had a positive loading on a factor which contained negative loading for total class gain in reading comprehension and English grammar, and a positive loading for total class gain in reading vocabulary.

## Table 5.1

## Teacher Talk

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Flanders, 1970<br>2nd - General<br>(15 tchrs)<br>Two semesters | | teacher talk<br>$r = .30$ |
| Flanders, 1970<br>4th - Social Studies<br>(16 tchrs)<br>Two weeks | | teacher talk<br>$r = .08$ |
| Flanders, 1970<br>6th - General<br>(30 tchrs)<br>Two semesters | | teacher talk<br>$r = .11$ |
| Flanders, 1970<br>7th - Social Studies<br>(15 tchrs)<br>Two weeks | | teacher talk<br>$r = .02$ |
| Flanders, 1970<br>8th - Math<br>(16 tchrs)<br>Two weeks | | teacher talk<br>$r = .45$ |
| Perkins, 1965<br>5th - General<br>(27 tchrs)<br>Two semesters | Factor II Teacher<br>Lecturer-Criticizer<br>+[a] reading vocabulary<br>- reading comprehension<br>- English grammar | Factor II Teacher<br>Lecturer-Criticizer<br>ns[a] arithmetic<br>ns spelling |

---

[a] + refers to positive loading on a factor containing this variable; ns refers to no loading on the factor, and - refers to a negative loading. Loadings for total class gain or loss in achievement not given.

Table 5.1 (continued)

Teacher Talk

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Sharp, 1966<br>High School-Biology<br>(31 tchrs)<br>Two semesters | | teacher talk<br>$r$ = .29 |
| Solomon et al., 1963<br>College evening school - American History<br>(24 tchrs)<br>One semester | | Factor 1: Control<br>proportion of teacher talk of total classroom speech (.92)[a]<br>$r$ = .19 (factual gain)<br>$r$ = .32 (comprehension gain) |
| Wright and Nuthall, 1970<br>3rd - Natural Science<br>(17 tchrs)<br>Three 10-minute lessons | | teacher talk<br>$r$ = -.09<br>teacher utterances<br>$r$ = .35 |

[a] Not all variables which loaded on this factor are given in this table. Only those variables relevant to teacher talk are presented.

The overall low, non-significant, but consistently positive correlations between teacher talk and student achievement appear to be different from the frequently voiced idea that teachers spend "too much" time talking in class. Although teacher talk as currently practiced does not appear to be significantly related to student achievement, the correlations are consistently positive.

"Teacher talk" is a rather gross category, and the investigators included this variable (and those following) as one of the many which they studied in any single investigation. The results suggest that there may be higher yields in focusing upon types of teacher talk rather than sheer amount, and examples of such variables are contained throughout Chapter Two (Affective Variables) and Chapter Three (Cognitive Variables).

## Student Talk

Frequency counts on student talk were used in the studies reported in Table 5.2. None of the results was significant or near-significant in any of the three studies.

It is rather surprising to find only three studies on this variable. Additional information on the relationship of student talk to student achievement could be obtained if the data from the five studies by Flanders (1970) were analyzed with this variable in mind, but the lack of even consistent correlations in the three studies in Table 5.2 does not induce much hope that additional analyses or studies would yield more fruitful results. Again, the non-significant and inconsistent findings in Table 5.2 are quite different from the "expert opinion" which stresses the importance of student talk.

Perhaps the lack of results for teacher talk or for student talk may be due to the grossness of these variables. It is because of such grossness that the results on these variables are presented last in this report; the division of teacher or student talk according to type appears much more promising, and such results are reported in Chapters Two, Three, and Four. In future studies even the variables in those studies might be refined and provision made for the inclusion of additional events in the category system. Such events might include the events preceding and following the talk, the context in which the talk takes place, and the specific subject being discussed. For the interested reader, expanded discussions on the development of new category systems are available elsewhere (see Biddle, 1968, Meux, 1968; Rosenshine, 1970; Rosenshine, in press).

Table 5.2

Student Talk

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Sharp, 1966<br>High School - Biology<br>(31 tchrs)<br>Two semesters | | student talk<br>$r = -.07$ |
| Soar, 1966<br>3rd thru 6th - General<br>(55 tchrs)<br>Two semesters | | Factor 2: Extended Student Talk[a]<br>sum of student talk (.89)[b]<br>extended student talk (.94)<br>Mdn. $r = .15$ |
| Wright and Nuthall 1966<br>3rd - Natural Science<br>(17 tchrs)<br>Three 10-minute lessons | | student talk<br>$r = .02$<br>extended student talk<br>$r = -.23$ |

[a] Not all factor variables are presented. Only those related to student talk are in this table.

[b] Factor loadings; not correlations.

## Teacher-Student Interaction

The amount of teacher-student *interaction* is a somewhat more focused variable than counts of teacher talk (Table 5.1) or student talk (Table 5.2). The variables in Table 5.3 refer to the number of teacher-student interchanges, frequency of teacher questions, or percentage of interactions which were classified as questions. The *type* of question or interaction is not considered in Table 5.3; those variables were presented and discussed in Chapter Three. Three of the studies listed in Table 5.3 were also cited in Chapter Three (Harris and Serwer, 1966; Harris et al., 1968; and Soar, 1966), because in those studies *both* frequency of interchanges and type of interchanges were categorized.

Significant results relating the frequency of teacher-student interaction and at least one measure of residual student achievement were obtained in four of the ten studies (Harris and Serwer, 1966; Soar, 1966; Wallen, 1st grade, 1966; Wallen, 3rd grade, 1966). Inspection of Table 5.3 will show that these significant results were not obtained on all criterion measures in the two studies by Wallen. Unfortunately, in four of the five studies by Flanders (1970), there is not even a trend or any consistency favoring a high frequency of teacher questions. This discrepancy between the studies by Flanders and those conducted by other investigators is puzzling, particularly because Soar and Wallen used Flanders' Interaction Analysis as their observational system.

The overall results are mixed. In four of the ten studies, significant results were obtained on at least one criterion measure. In five studies which yielded non-significant results (Harris et al., 1968; Flanders, 2nd grade, 1970; Flanders, 4th grade, 1970; Flanders, 7th grade, 1970), the correlations were both small and erratic (*rs* = -.19 to .11).[1]

One additional study of amount of teacher-student interaction--not included in the tables because of small sample size--also yielded mixed results. Lahaderne (1967) found significant adjusted correlations (*rs* = .3 to .5) between the frequency of instructional interactions and various standardized measures of pupil achievement, but pupils were the unit of analysis in this observational study of four sixth-grade classrooms.

[1] [The reader should not that these results on teacher-student interaction or frequency of questions represent a shift from the results reported in an earlier summary of research in this area (Rosenshine, 1969). The earlier review did not include the five studies by Flanders (1970).]

Table 5.3

Teacher-Pupil Interactions (Frequency of Questions)

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Harris and Serwer, 1966<br>1st - Reading<br>(48 tchrs)<br>Two semesters | r (observer counting)<br>total interchanges<br>med. r = .32* | |
| Harris et al., 1968<br>2nd - Reading<br>(38 tchrs )<br>Two semesters | | r (observer counting)<br>total interchanges<br>med. r = -.03 |
| Flanders, 1970<br>2nd - General<br>(15 tchrs)<br>Two semesters | | r (observer counting)<br>percentage of questions<br>r = .07 |
| Flanders, 1970<br>4th - Social Studies<br>(16 tchrs)<br>Two weeks | | r (observer counting)<br>percentage of questions<br>r = -.19 |
| Flanders, 1970<br>6th - General<br>(30 tchrs)<br>Two semesters | | r (observer counting)<br>percentage of questions<br>r = .11 |
| Flanders, 1970<br>7th - Social Studies<br>(15 tchrs)<br>Two weeks | | r (observer counting)<br>percentage of questions<br>r = -.05 |

*p $<$ .05

Table 5.3 (continued)

Teacher-Pupil Interactions

| Investigator | Significant Results | Non-significant Results |
|---|---|---|
| Flanders, 1970<br>8th - Math<br>(16 tchrs)<br>Two weeks | | r (observer counting)<br>percentage of questions<br>r = .44[a] |
| Soar, 1966<br>3rd thru 6th - General<br>(55 tchrs)<br>Two semesters | r (observer counting)<br>Factor 3: Extended Discourse<br>inquiry/drill ratio (.60)<br>drill (-.81)<br>med. r = .28* | |
| Wallen, 1966<br>1st - General<br>(36 tchrs)<br>Two semesters | r (observer counting)<br>frequency of questions<br>med. r = .44*<br>r (observer counting)<br>percentage of teacher asking questions<br>r = .40* (reading vocabulary) | r (observer counting)<br>percentage of teacher asking questions<br>med. r = .32[a] |
| Wallen, 1966<br>3rd - General<br>(40 tchrs)<br>Two semesters | r (observer counting)<br>percentage of asking questions<br>r = 36* (arithmetic) | r (observer counting)<br>freq. of questions<br>med. r = .13<br>percentage of asking questions<br>med. r = .12 |

[a] $p < .10$

* $p < .05$

## Summary

Of the three variable reviewed in this chapter, two have some promise. Teacher talk yielded highly consistent, but non-significant correlations with residual gain scores. In the ten studies of the amount of teacher-student interaction, significant results were obtained in four, and low and erratic correlations in five others.

One might expect that the teacher-student interaction would serve to arouse attention, and the importance of attending behaviors and internal rehearsal has been demonstrated in laboratory studies using school age children and meaningful material (see Rothkof, 1966; Anderson, 1969; Travers et al., 1964). However, in these laboratory studies, the instructional materials were constant -- only the attention arousing procedures were varied. In normal instruction, many cognitive and affective activities vary from classroom to classroom, and some of these variables are covered in Chapters Two and Three. Given such variation, we might well expect that such gross variables as those reviewed in this chapter would have little relationship to student achievement.

# CHAPTER SIX

# FUTURE RESEARCH ON TEACHING BEHAVIORS[1]

In this final section the major findings to date are summarized. The emphasis, however, is upon suggestions for future research in this area. The major section is on future process-product studies because of the importance which many researchers and educators give to knowledge developed in such settings. Many of the ideas for improved process-product research also apply to the development of experimental classroom studies, an area which has been badly neglected.

## Summary of Results

In Chapter Two, on teacher affective behaviors, process-product relationships were reviewed in six categories of teacher behavior: criticism and control, non-verbal approval, praise, use of student ideas, indirectness, and indirect/direct ratios. In none of these categories were there significant results for more than half the studies, but there were consistent positive trends for use of student ideas, indirectness, and indirect/direct ratios, and a consistent negative trend for criticism.

Although there were too few studies to warrant confident conclusions, it appeared that specific types of praise, use of student ideas, and criticism yielded higher correlations than any entire category. The specific types of praise or use of student ideas which yielded significant results in any study are too various to permit comparison. Some comparison is possible on types of criticism. In no study was mild criticism (e.g., telling a student that his answer was wrong) negatively related to achievement; however, stronger forms of criticism such as "hostile or strong disapproval" (Hunter, 1968) and disapproval by shaming (Spaulding, 1965) yielded significant negative correlations.

Fewer studies were found which focused specifically upon cognitive variables (Chapter 3). The strongest finding in this area was the lack of a significant linear relationship between the frequency of use of any type of question and student achievement. When questions or types of discourse were classified into more than two categories, significant results were obtained in all three studies, but the category systems were too diverse to permit synthesis of the results. There were suggestions that responding to student answers by asking further questions (probing), providing structuring comments at different points in a lesson, and focusing clearly upon intellectual activities are behaviors significantly related to student achievement, but, again, too few studies were completed to permit statements about the strength and consistency of these findings.

---

[1] Many of the ideas from previous papers (Rosenshine, 1969; Rosenshine, 1970a, b, c, d; and Rosenshine and Furst, 1970) have been included, expanded, and/or repeated in this section. Many of the ideas in this chapter are those of Norma Furst, or were developed in our conversations while writing our 1970 paper.

Variation in specific classroom activities or in types of cognitive interactions (Chapter 4) appeared to be a clear indicator of student achievement, although the number of studies in this area was small. There were consistently positive but non-significant correlations between the amount of teacher talk and student achievement (Chapter 5), but consistent results were not obtained for student talk, nor for the frequency of teacher-student interactions.

Teaching strategies for student achievement. There are suggestions in this research that the most effective teachers do not merely emit a specified number of approving statements or types of questions; rather, they may use certain behaviors and avoid others in order to achieve particular cognitive ends. For example, the most effective teachers studies by Conners and Eisenberg (1966) emphasized interchanges involving intellectual content, and avoided interchanges involving property and materials; the most successful teachers studied by Spaulding (1965) approved pupil responses which gave interpretation and judgment, but they asked few questions related to pupils' interests, and few open-ended questions; the successful teachers studied by Soar (1966) and Furst (1967) gave short structuring lectures before they asked questions; the successful teachers studied by Solomon et al. (1963) emphasized both factual and interpretive questions; and in three investigations, the most successful teachers responded to a pupil answer by "probing," or asking the pupil or the class to elaborate and clarify what was said (Spaulding, 1965; Soar, 1966; Fortune, 1967).

In each case, the teacher may have chosen the effective behavior because he thought the behavior would advance the attainment of specific cognitive ends. A moderate amount of structuring before a question may have been used because such structuring appeared to improve the quality of student answers. It was not student participation alone that the teacher sought, but a certain quality of response. Only selected student responses were approved because these were the ones the teacher wished to encourage; pupils were asked to elaborate and extend their answers because such student behavior moved the class discussion toward certain ends that the teacher had in mind. At the same time, the teacher avoided behaviors which did not contribute toward cognitive ends, such as emphasis upon property and materials, questions related to students' interests, or excessive criticism. If ends-in-view are a critical component of effective teaching, then we should expect that increasing only the teachers' use of specific behaviors would have minimal effects.

One additional general suggestion can be based upon the research on cognitive interchanges. After the primary grades, single cognitive behaviors are not significant correlates. Rather, the overall pattern of behaviors is more important. Such a pattern includes the use of a variety of questions, moderate amounts of structure, lesser amounts of drill, and frequent requests for the pupil to elaborate his answer.

## Future Process-Product Research

Years ago, Ackerman (1954) and Morsh and Wilder (1954) called for research on teaching which would employ systematic observation of specific teaching behaviors and would correlate these behaviors with measures of pupil achievement. Such research, they suggested, would be more productive than the previous studies which had utilized general rating scales and measures of teacher personality and characteristics as independent variables.

When the 35 studies reviewed here--which do relate systematically observed behaviors to measures of pupil achievement--are contrasted with the previous studies which compared teacher characteristics and personality to measures of pupil achievement, the comparison does not overwhelmingly favor the more systematic approach; the results are not as clear or conclusive as Morsh, Wilder, and Ackerman expected. Their expectation that the counting of relatively objective teaching behaviors would yield consistent, significant correlations with student achievement certainly has not yet been fulfilled. Indeed, the most promsing results have been obtained in studies in which teacher behavior was described on rating scales by classroom observers (see Appendix). The results obtained on variables such as clarity, enthusiasm, and task-orientation appear very promising.

After 10 years of process-product research, 35 studies, and mixed results, some researchers would claim that such correlational research will not be productive in the future. Because of the limited research, and because of the methodological problems which may exist in most of these studies, any judgment on the worth of this research would be premature. However, before any conclusion is reached, perhaps there should be at least a second generation of this research incorporating some of the suggestions presented below. These suggestions cover four major areas: selection of variables, procedures for coding classroom events, administrative design of the studies, and statistical procedures for analyzing the results. Some topics discussed in each of these areas are applicable to more than one area.

## Selection of Variables

Four suggestions are offered for the selection of variables in future process-product studies: (1) use of variables available in existing observational category systems and rating systems; (2) use of a greater variety of variables, such as more comprehensive cognitive variables (e.g., multiple classification of questions, use of varied activities, and similar variables cited in the above review); (3) use of more variables developed from laboratory studies; and (4) use of high-inference and low-inference variables together in the same investigation.

It is not particularly difficult to select a large number of variables for use in an observational category system; at least 200 systems have been developed. Although many of the variables overlap or are duplicated in different systems, a large pool of distinct variables has been developed. Not only is there a variety of variables, but there is also a variety of units of measure and contexts in which the classroom events occur.

It is sad that although many observational category systems have been developed, so few have been used to relate frequencies of the variables to measures of student achievement (or any ctiterion measures). Only 35 process-product studies have been found, and in 21 of these, Interaction Analysis (or a variation of this system) was the observational system. Almost all of the 80 observational systems anthologized by Simon and Boyer (1970) have been used primarily to collect descriptive data on classroom processes; no more than 10 have been used in a process-product study.

Most of the process-product studies which have been discussed have focused on affective variables. In the introduction to Chapter Three, the difficulties of coding cognitive variables and the lack of research in this area were discussed in greater detail. Some investigators have developed classroom observational category systems which focus on cognitive interactions. Unfortunately, few investigators have used these systems to attempt to determine which of the cognitive variables are related to measures of student achievement. More research on cognitive variables seems warranted. Promising but insufficiently researched variables include multiple classification of questions, probing responses to student answers, variation of activities and of the cognitive level of the discourse, and use of structuring statements.

There are also several cognitive variables which apparently have not yet been used in ovservational category systems. Variables such as the relevance of the materials to the ability of the class, or the amount of time a teacher spends preparing a class for future classwork have not appeared in the systematic observational systems because these variables are extremely difficult to quantify. Indeed, most of the cognitive variables which are discussed in educational psychology textbooks have not been included in these category systems.

Selecting Variables from Laboratory Research. There has been considerable study in Laboratory-type settings of meaningful human learning and the effects of different types of instructional materials upon achievement. But there is little overlap between the variables developed for use in classroom observational studies and the variables being investigated in laboratory research and in research on instructional materials. For example, in one review of specific treatment variables associated with instructional materials (Popham, 1969), the major headings were: organizers, relevant practice, knowledge of results, promoting learner interests, prompts, sequencing, and pacing. An anthology of research reports on meaningful human learning in laboratory settings (Anderson et al., 1969) included the following titles as section headings: prompting and fading techniques, the student response, reinforcement and feedback, facilitation of concept learning, and organization and sequence. By contrast, in the current review of classroom observational studies the variables included "indirectness," "teacher talk," "multiple classification of questions," and "variation."

This lack of common variables between laboratory and classroom research may have occurred because studies of "instruction" in classrooms have focused on instruction mediated by a teacher. In effect, two separate disciplines are being developed to study meaningful human learning. One contains a minimum and the other a maximum of verbal interaction. Although there is some overlap between the two disciplines in areas such as reinforcement and feedback, there has been little attempt to assimilate one with the other. Occasionally, bridges are built. Nuthall (1968) used programmed materials to investigate the effects of classroom instructional strategies identified by Smith, Meux, et al. (1967), and the study by Worthen (1968a, 1968b) was explicitly designed to test whether the laboratory studies on discovery learning could be replicated in a natural classroom setting. Perhaps more such interaction will develop, and variables developed in the laboratory will be applied to classroom research and vice versa. For example, I would hope that many of the ideas on "test-like events" could be applied to correlational and experimental classroom research.

Employing High Inference and Low Inference Measures. Although this review focuses upon the results of studies in which low inference measures were used, a glance through the appendix will show that despite frequent comments minimizing the usefulness of rating scales, many of the strongest results were obtained through the use of ratings of specific teacher behaviors made by students or outside observers.

The advantage of rating scales may be that they allow the rater to process a large number of cues before he makes a decision on the teacher's "clarity," "enthusiasm," or "task oriented behavior." In contrast, someone using an observational category system is unable to perform such processing because he is required to record specific behaviors.

The fact that rating systems appear valuable for identifying gross or high inference teacher behaviors which are related to student achievement has apparently been overlooked during the recent period of emphasis upon observational category systems. But there is no need to decide whether category systems or rating systems are more useful. The optimal strategy would be to employ both types of observational systems in future studies of teaching, and to determine which specific low inference variables best describe the items on rating scales that are most predictive of student achievement. Therefore, the more consistent findings from studies in which rating scales were used might also serve as sources of variables to be used in future observational category systems.

## Procedures for Coding Classroom Events

Recent developments in observational category systems might be classified under four overlapping areas: scope of the behavior included in a category, development of methods for identifying the sequence of events, development of methods for coding concurrent events, and development of varied analytic units.

*Subdivision of variables*. As this research has continued, investigators have begun to subdivide relatively large categories such as praise or use of student ideas into smaller, more specific behaviors, and some of the smaller variables have had significant correlations with student achievement. For example, when Wallen (1966) and Perkins (1965) divided control and criticism into types, they both found significant negative relationships for personal control, but not for academic control.

The process of subdividing larger categories into smaller ones has been labeled "subscripting" (Flanders, 1970). A number of suggestions for the subscripting of categories might be derived from the research to date. The major variables (or categories) discussed in this section are praise, use of student ideas, criticism, and questions.

The generally non-significant results obtained when frequencies of use of praise were tabulated may be due to the grossness of this category. The results of a number of studies suggest that praise can be subdivided into four forms: (1) mild praise which indicates the correctness of an answer, (2) strong praise, such as saying, "Great!," (3) extended praise in which a reason for the praise is given, and (4) extended praise in which the praise is repeated in different words. The two forms of extended praise may be similar to teacher use of pupil ideas in that both behaviors indicate that the teacher has listened carefully to the pupil's comments. It would also be of interest to see whether mild praise which contains a reason for the praise differs in effectiveness from strong praise such as "Great!" The research by Spaulding (1965) suggests that additional subscripting could profitably be used to record the specific behavior being praised, such as pupil interpretation of ideas, pupil knowledge of expected answer, or pupil attention to task.

*Use of Pupil Ideas*. The variable labeled "teacher use of pupil ideas" has a good history as a correlate of achievement. Higher frequencies of use of this variable yielded relatively moderate, consistent correlations with achievement, and this variable forms a major part of the i/d ratio. Yet it is a rather gross category which, because of its good record, merits detailed study. Various approaches are possible.

Within Category 3, additional information might be obtained by "subscripting" teacher responses. Flanders (1970) has noted five such behaviors: repetition, modification, application, comparison, and summary of pupil ideas. I would be interested in the results obtained when these more specific variables are studied singly or in combination.

It would also be interesting to find alternative procedures to achieve the same cognitive results achieved by the teacher's acceptance and use of student ideas. One procedure would be for the teacher to ask other students to summarize, compare, or elaborate what one student said. Such repetition might involve the affective components of giving publicity and indicating that someone has listened to the student; it may also have cognitive elements because it provides reiteration and clarification of key points. In addition, requiring such student behaviors and approving their occurance may facilitate the student's implicit rehearsaland practice of the major cognitive processes involved in the lesson.

Additional alternatives for expansion and use of student ideas have been discussed above under the heading "probing responses" in the section of cognitive results (Chapter3). There is some indication that several behaviors which appear to resemble use of student ideas are uncorrelated with one another. Thus, the two significant positive behaviors identified by Spaulding (1965) loaded on two different factors: teacher approval of pupils' interpretations, and teacher disapproval by eliciting clarification in a non-threatening way. In the study by Soar, there were positive results for the teacher's encouragement of pupils' elaboration and generalization, although such behaviors were uncorrelated with teacher use of behaviors in Category 3. The existence of these similar but uncorrelated behaviors further indicates the complexity of this particular area, and the difficulty of separating the cognitive and affective components.

Because of these varied results, there is a need for future investigations which subscript the behaviors within Category 3 and which include behaviors that superficially appear to resemble use of pupil ideas. Once such investigations are complete, we should have more specific knowledge about the number of factors which reside within Category 3 and the alternative forms of this behavior, and about which of these factors are consistent correlates of certain pupil product measures.

Another variable that could be subscripted is criticism, which could be divided into two types: extended criticism, and criticism for which a reason is given. There is also a need to separate criticism and directions concerning academic activities from criticism and directions concerning personal control.

A large number of category systems have been developed for subscripting the cognitive aspects of instruction (see Simon and Boyer, 1970). Unfortunately, most investigators used these systems primarily to describe classroom discourse, and few have studied the relationships between cognitive behaviors and student achievement. Most of the cognitive systems have used the classification scheme of the Taxonomy of Educational Objectives, Handbook I: Cognitive Domain (Bloom, 1956) or the three factor scheme devised by Guilford (1967), although other investigators have developed unique systems based upon classroom observation(e.g., Smith et al., 1962). The research results on cognitive variables (Chapter Three) suggest that there is an advantage to using category systems which divide questions (or discourse) into more than two types. The available cognitive observational category systems await use.

The theme of these suggestions for future research has been the value of breaking categories into smaller units of behavior, and studying the relationship of frequencies of these behaviors with achievement. Such suggestions have been supported by reference to research reported in the preceding chapters. Of course, it is possible that the units can become too small for use. Such concerns can be tested empirically; at present we need to understand the smaller variables better.

Subscripting and Inter-investigation Reliability. The use of subscripting in future studies may alleviate some of the problems which occur when different investigators include different behaviors in the same general category. In Chapter One, it was noted that some investigators included teacher repetition of student ideas as part of praise, whileothers included the same behavior as part of use of student ideas. The use of subscripts allows an investigator to make a separate count of behaviors as a subscript under praise or use of student ideas. Once such a special category (or subscript) is created, the frequency of behaviors in this category could be analyzed separately, or combined with either praise or use of student ideas. If such procedures are followed, and if the original counts are presented in the final report, then the investigator or reviewers of the research can reanalyze the data according to trends in results from other studies. Currently, without subscripting of categories, the problems of inter-investigation reliability may hinder selection of the most appropriate method of analyzing the data.

The use of subscripts is not without problems. Increasing the use of subscripts will probably make the category system difficult to use and unwieldy to analyze. Many users have been attracted to the Interaction Analysis system because the 10 category system is easy to teach and use. A more complex category system may be more useful for research, but it may not be the best instrument for training teachers or for helping them to observe their classroom behavior. The use of subscripts also raises the question of the optimal size of these smaller units. It is possible to identify 10 forms of "silence or confusion," and 10 types of questions, but we do not know whether the results would be worth the extra work. Empirical study of the advantage of increasing the number of subscripts is difficult because the number of subscripts which can be created far exceeds the number of teachers in the usual sample studied in process-product studies.

Coding Concurrent Events. With the exception of the system used by Spaulding (1965), all the observational category systems used in the above studies might be classified as one-dimensional or one-factor systems. That is, each behavior is coded only in terms of its frequency, and concurrent events are not included. For example, instances of teacher praise are recorded, but the systems do not provide for recording what action of statement is praised, or the content, level of conceptualization, or topic of the action or statement. Because of this problem, two teachers may be coded as having identical percents of evaluative questions, yet one teacher may have been discussing use of a microscope and the other may have been discussing the decoration of a bulletin board. Similarly, teacher praise regarding student knowledge may be a different variable from praise of student persistency. Student persistency itself can differ in context; in one class the students may be attempting persistently to sound out new words, and in another class students may be drawing pictures persistently.

Two investigators have independently developed two similar approaches for coding concurrent events. Gallagher (1970) has labeled his the Topic Classification System, and each "topic" of classroom discourse is coded three ways: according to emphasis upon skills or content; the level of conceptualization (e.g., data level, generalization level); and the logical style used by the teacher (e.g., description, generalization, expansion). Flanders (1970) has labeled his approach "multiple coding," and he codes each category (mainly affective categories) according to the type of move (see Bellack et al., 1966) and the cognitive level (see Taba, 1964). The advantage of both these approaches is that the coding system provides more information on the context of classroom events such as types of questions or types of teacher responses.

There are numerous contextual variables which could be included in a multiple coding system. Some might be: student attention to task, the specific area of content being considered, the accuracy of the student or teacher statement, and the cognitive level of the content. Unfortunately, no one has used multiple coding as part of his analysis of the data. In the descriptive research reported by Gallagher (1970), in which he used the Topic Classification System, separate results were reported for skill versus content, level of conceptualization, and logical style. In the study by Bellack et al. (1966), separate analyses were made of the type of pedogogical moves (e.g., teacher solicits, student responds), the thought process occurring (e.g., explaining, evaluating), and the substantive area of the materials being studied (i.e., the topic).

The concepts of subscripting and multiple coding are quite new, and the fine distinctions between each approach are yet to be made. At present, the subdivision of behaviors into smaller units appears to typify subscripting; the inclusion of additional contextual variables appears to typify multiple coding. In practice, the two innovations can overlap. For example, in addition to subdiving teacher directions into managerial directions and academic directions, one could further subdivide them into thought processes, type of pedogogical move, or topic being considered. Such subdividing appears similar to multiple coding. Perhaps the best distinction which can be offered at present is that subscripting focuses primarily upon subdivision of the behavior, whereas multiple coding focuses upon the context of the event. As an investigator modifies his system in order to obtain more information on the behavior or the event, the two procedures appear to coincide. One example of a category system which appears to be an example of either subscripting or multiple coding is the one used by Spaulding (1965). In Spaulding's system, teacher behaviors were classified according to their (a) major type (e.g., approval, instructional), (b) source of authority, (c) number of class members included in the statement, (d) amount of attention the class gives to the statement, (e) tone of voice, (f) technique used, and (g) topic.

The need to obtain more complete information on classroom behavior includes the need for data on the sequence of events. Although investigators have counted the frequency of teacher behaviors categorized as rewarding and punishing, little attention has been paid to specifying the event which was rewarded or punished. Similarly, little distinction has been made between the use of structuring statements at the start of a series of lessons and the use of structuring statements within a lesson. Investigators have tended to use only the total frequency of structuring statements in their statistical analyses. Similarly, no one has investigated the effective difference between asking a broad question at the start of a lesson and asking a broad question at the end of a lesson.

Interaction Analysis is one category system which provides some data on the sequence of behavior because events are recorded as diads; thus, by inspecting the matrix, an investigator can determine the dominant pattern of interaction. However, such a procedure is relatively gross and cannot be used to answer the questions posed above. Many investigators are attempting to develop methods for preserving sequence, but I did not find any report of a process-product study in which such procedures were used. These suggestions that contextual behaviors and the sequencing of behaviors be considered also imply that it may be fruitful to study teaching as a strategy. But although terms such as teaching strategy and teaching style are commonly used, investigators have not yet been able to define these terms using specific, denotable behaviors.

Analytic Units. Early investigators have coded classroom events according to their duration. The Flanders Interaction Analysis System with its "three second rule" is an example of the use of time as the primary analytic unit. Other investigators have attempted to develop cognitive units within which the frequency of events is recorded. These investigators have developed complex units such as a "move" (Bellack, 1965), a "venture" (Smith, 1964, 1967), or a "topic" (Gallagher, 1968). These new analytic units are then coded as to their dominant cognitive process, the types of questions which occur within them, or teacher affective behavior. Although analytic units such as these are difficult to use, it does not follow automatically that other units that are easier to use--such as time or lines on a transcript--should be substituted for them. Whether a cognitive unit, a time unit, or a combination of the two is the most appropriate unit for studying classroom interaction is an empirical question which has received too little study. Perhaps the question of the appropriate analytic unit will be studied in the second generation of observational classroom studies.

## Design of Process-Product Studies

Given a set of variables to represent important aspects of classroom instruction, and given a set of procedures to record the frequency, context, and sequence of these behaviors, the next problem is the design of an appropriate means to relate the observed behavior to the measured outcome.

The most frequently used design is one in which a pretest (or pretests) is given at the start of the semester, teacher and student behavior is sampled during a school year, and a posttest is administered at the end of the year. Such "long term" studies have been criticized because in such situations there may not be an appropriate match between the curriculum materials, the teacher's aims and behaviors, and the criterion instruments. For instance, the question has been asked (G. Nuthall, personal communication), if one group of teachers is teaching skills A, B, and C well, and another group is teaching skills D, E, and F poorly, what will be shown if their classes are tested on skills X, Y, and Z?

Such criticism is particularly cogent if standardized achievement tests are used as criterion measures (Flanders, 1970). Such tests may be inappropriate measures of the influence of the teacher's behavior if the items on the tests are not relevant to the materials or skills taught in the classroom. Teachers may not be interested in standardized achievement tests (Jackson, 1968). In many studies, these tests may have measured the aptitude of the learner or the pressure for academic achievement in the home rather than the influence of the teacher.

Currently, we may be faced with the problem of teachers teaching for various goals, few or none of which may be related to the criterion tests, and researchers trying to see which teacher behaviors are related to goals that neither the teacher nor the students perceives. However, it is possible to devise alternative designs in which there is more congruence among the curriculum, the teacher's behaviors, and the criterion instruments. These new designs, to be discussed below, focus on increasing the investigator's control over the teaching situation. The paradox is that the new situation may not represent naturally occurring teaching as it presently exists. Given the diverse goals of teachers, curriculum developers, students, and test developers, we question whether adequate designs can be developed to study the relationship between teacher behavior and student achievement in the typical, uncontrolled, classroom situation.

Possible Modifications in Design. Some of the above problems might be alleviated if we were to study teacher behavior for a shorter time, such as instructional periods ranging from 15 minutes to 10 one-hour, daily lessons (short-term studies). When the instructional period is short, we can specify the criterion measures, control the instructional content by providing the materials, give the teacher some examples of the criterion measures so that he can focus the instruction upon relevant material, observe the instructional period, and record it on audiotype or videotape. Studies employing this design offer the promise of focusing attention upon specific aspects of the teacher's role, such as the ability to explain new material; investigators will not have to contend with other "noise" such as the teacher's managerial and disciplinary function.

Such a concern for specificity and control has led to a number of studies (e.g., Furst, 1967; Flanders, 1965; Rosenshine, 1968; Wright and Nuthall, 1970), and the results of these studies are included in the preceding review. Surprisingly, these studies have not yielded a more significant results than those obtained in the long-term studies, nor was there any different pattern of findings.

The lack of stronger results in the short-term studies leads to two suggestions. One concerns the coding of the instruction for its relevance to the criterion tests, and will be discussed in the next section. The second suggestion is that further efforts be made to stabilize the behaviors of the teacher before the study is begun so that there is greater congruence between the criterion test and the teacher behaviors.

In the reported short-term studies, even though the teachers were given specific instructional materials and told the type of questions that would be on the criterion test, the use of content and cognitive processes was not controlled. As a result, in one study (Bellack et al., 1966), although all teachers and their students were given the identical pamphlet, there was wide variation among classes in the content covered and in the type of cognitive processes which the teacher called for in the teacher-student interchanges. In another short-term study (Wright and Nuthall, 1970), the teachers were given outlines of the material to cover each day and were told that the test would be factual. Yet some teachers asked open-ended questions, or responded to student answers with a further question designed to raise the cognitive level of the student response. In this study the percentage of open-ended and reciprocal questions was negatively (though not significantly) related to achievement. The authors concluded that although the teachers may have been attempting to teach thinking skills through such questions, such behavior was inappropriate for the criteria of the study.

In the two examples above, even though there was a considerable control built into the design, there were still wide variations in the behaviors of teachers. In the context of these studies, such variation represents "noise" because the behaviors were inappropriate to the criterion measures. We do not know what correlations between teacher behavior and student achievement would have been obtained if the teachers had been trained in criterion-specific behaviors before they began their instruction.

Perhaps the next step in increasing control in process-product studies would be to stabilize the teacher's behavior through training so that the observed behavior is a more accurate reflection of the teacher's intention and/or the intentions of those who prepared the instructional material. Curriculum developers and teacher educators would have to work together on this problem. Without such cooperative work we may continue to have curriculum experts developing instructional packages without clearly specifying teacher behaviors, and teacher educators training teachers in teaching skills without clearly specifying the instructional situations in which they will be used.

## Methods of Analysis

Five relatively distinct issues concerning the analysis of data from process-product studies are discussed in this section, in the hope that future investigators will contribute to the resolution of these issues or consider these problems in the design of future studies. The five issues are: controlling for opportunity to learn, types of statistical analyses, selection of variables for analysis, methods for reanalyzing existing studies, and the assumption that there is one set of "good teaching" procedures.

Opportunity to Learn. In the previous section in which the major results of process-product studies were summarized, the variable, "student opportunity to learn" the criterion material, was cited as a consistent and significant correlate of student achievement. Such a variable has not been sufficiently considered in the analysis of process-product studies; in almost all studies no measure was taken of student opportunity to learn, and consequently classes were treated as if they all had had equal opportunity to learn. One procedure for assessing opportunity to learn is that used in the international study (Husen, 1967), whereby teachers estimated the percentage of students who had had an opportunity to learn material of the type exemplified by each test item. Such a procedure could be applied to studies in which standardized achievement tests or special curriculum tests are used as the criterion. For example, a teacher could be shown the questions which follow a reading selection and asked whether the students in his class had an opportunity to learn the processes necessary to answer such questions. Similar procedures could be used for most areas such as arithmetic concepts and problem solving, map skills, or application of biological laboratory principles. When short-term studies are conducted, the transcripts or tape recordings of the class sessions could be inspected to determine whether the criterion material was indeed covered (see Rosenshine, 1968; Shutes, 1969).

The data on opportunity to learn have been used as a covariate to adjust the posttest scores further before searching for teacher behaviors related to the adjusted posttest measure (Rosenshine, 1968). But the data could also be used as a correlate of achievement, as a relevant teacher behavior contributing to student achievement (Shutes, 1969).

Statistical Analyses. The difficulties in obtaining reliable measures of student achievement, and the various methods of analyzing the data have been discussed in Chapter One. The major conclusion of that discussion was that the type of statistical procedure was related to the purpose of the analyses. If a major purpose of process product studies is to identify promising variables for use in future experimental studies, it does not seem appropriate for investigators to limit themselves to any given level of statistical significance or to one set of statistical procedures. Rather, a variety of procedures should be used to identify promising variables.

One general statistical procedure which yielded interesting results and hypotheses for future study was non-linear analysis. In some of the studies reported in Chapters 2, 3, and 5, significant results were obtained when non-linear analyses were used. Such procedures seemed particularly useful in studying cognitive variables, although Soar (1969) has shown significant non-linear relationships in the study of affective variables. Such results suggest that future investigators should give more attention to non-linear analyses. There may also be benefits from reanalyzing existing studies using some of the non-linear procedures specified in the above chapters; at a minimum the scatter plots from existing studies ought to be studied. One may find that although linear correlations were not significant, there were significant differences between teachers who were at the extremes on a variable or on a criterion measure. Identification of the characteristics and behaviors of such teachers might be useful in designing future experimental studies.

Selecting variables for analysis. If one compares the number and variety of observational category systems to the number and variety of variables which have been studied in process-product studies, one concludes that it must be easier to devise complex category systems than it is to analyze their results statistically. An investigator who uses subscripts, multiple coding, and/or a matrix to code classroom behavior obtains an extremely large number of variables which can be correlated with achievement. Consider the relatively simple 10 category system of Interaction Analysis. Hundreds of variables can be drawn from the 100 cell matrix by selecting individual cell frequencies, combining cell frequencies, or forming ratios of one set of cells to the other. If the investigator has expanded the IA system by subscripting all or some of the categories, he can easily obtain a 42 X 42 cell matrix which can yield thousands of variables for analysis. The same problem of selecting variables for analysis occurs in other systems which have used multiple coding, such as those developed by Bellack et al. (1966) and

Gallagher (1970). A very large number of variables could be selected from such systems for statistical analysis.

In practice, investigators have made apriori selection of individual cells, combinations of cells, ratios of cells, or composites formed from combinations of cells and used these in the statistical analyses. The number of variables which they submit to statistical analysis tends to be very conservative in comparison to the number of variables they could select. For example, although each used the 10-category IA matrix to code classroom behavior, Flanders (1970) selected only 15 variables, and Soar (1966) selected only 39 variables for statistical analysis.

One can appreciate such caution. As one increases the number of variables (particularly when the number of classrooms is relatively small), the risk of spurious significant results increases. Most investigators have decided to use a cautious approach. However, I do not believe that such caution is warranted for at least two reasons. First, the problems of doing research in natural settings are so large that we can expect confounded results as a matter of course. Limiting the analyses to a few variables does not reduce the logical and statistical problems of coding behavior and obtaining residual gain scores on specific tests. The best solution to such problems appears to be replication. Whether we obtain significant results in a single study is not as important as whether we obtain consistent results across a series of studies. If replication is the important end, then we need not be so concerned about "false-positives" because these will fall out across the replications. Such caution is also unwarranted because our primary end is not obtaining a set of teacher behavior variables which will predict class mean residual gain. Rather, our end is the improvement of instruction. The importance of the correlations we obtain, even those which are consistently significant acorss a number of studies, will be best tested in experimental studies (see below).

One unanticipated consequence of conservatism in selecting variables for analysis is that in the process the investigators throw away potentially useful data. For example, when an i/d ratio is used, the data available from 50 cells are reduced to a single variable for analysis. Whether the i/d variable is a stronger and more consistent predictor of achievement than the correlation which could be obtained from using other combinations of the 50 cells is an empirical question which has been neglected often in the analyses. The same problem occurs when investigators form composites of different sets of variables (e.g., Furst, 1967; Powell, 1968) without first (or also) determining the relationship between the individual variables and student residual achievement gain. Although it may be true that an i/d ratio or various composites are better predictors of achievement than any of the specific cell components, the possibility has not been tested empirically.

One solution to the problem of losing potentially useful information by combining a number of variables into ratios, clusters, or composites might by to conduct a two step analysis. In the first step, the investigator could develop his hypotheses and parsimoniously select a limited number of variables for statistical analysis. In the seond step, hundreds of variables could be formed from the data and subjected to analysis. In this second step, an investigator could use each of the 50 cells in the i/d ratio, combinations of cells from an IA matrix (or from an expanded, subscripted IA matrix), each of the cells and combinations of cells from Category 3 (use of student ideas), and any variety of measures of indirectness, directness, or ratios of the two. In studies in which composites were formed (e.g., Furst, 1967; Powell, 1968), each variable in the composite could be studied separately. The primary question in such a post hoc fishing expedition would be whether any of these new variables predicts students achievement as well as or better than those variables chosen originally. Such post hoc analyses could be conducted by the original investigator, and, if sufficient data were presented in the complete report, other investigators could perform these analyses. If the post hoc analyses revealed that certain variables were better predictors than those origninally selected, then the potency of these new findings could be checked by reanalyzing the data from another study.

There are several variables which might be chosen for such post hoc analyses. Because of the publicity which has been given to the behavior "use of student ideas," it would be interesting to know how well single cells, or combinations of cells within Category 3, correlate with student achievement compared to the correlation yielded by using the column total. We would be most interested in knowing whether frequencies in Category 3 are better predictors whentaken by themselves, or whether prediction is improved when they are used as part of an i/d ratio. Similar post hoc analyses could be applied to the cell and column frequencies in Category 6 (giving directions) and Category 7 (teacher use of criticism).

I performed two small post hoc analyses in preparing this review. In her initial report, Furst (1967) developed a composite consisting of three variables: the ratio of extended indirect to extended direct teacher behavior (see Figure 1.1), the i/d ratio for rows 8 and 9, and extended student talk (student talk lasting more than three seconds). A post hoc analysis showed that although the extended i/d ratio and the i/d ratio for rows 8 and 9 were significant variables by themselves, the variable extended student talk added nothing to the composite. Without such a post hoc analysis, I might have included extended student talk as part of the significant findings in this review.

A similar post hoc analysis was performed on the data provided by Hunter (1968). In this analysis the question was whether colomn totals or indirect/direct ratios yielded higher correlations when student achievement: the correlations for various i/d ratios were higher than the correlations of column totals (i.e., specific behaviors) considered separately.

Criteria for stratifying teachers. One common practice in studies in which Interaction Analysis is used has been to split the sample into two, labeling one group direct and the other group indirect teachers. Unfortunately, the degree of teacher indirectness may vary from study to study, and a teacher classified as indirect in one sample might be classified as direct in another. More useful results might be obtained if teachers were stratified according to their use of certain behaviors. One type of stratification which would be particularly useful in studies in which an i/d ratio is used might be "i/d about 2," i/d above 1," and so on. This stratification would facilitate more precise interpretations of the relationship between levels of a variable and achievement and would allow those involved in teacher education to describe "indirect" and "direct" teaching in more specific terms. Such a suggestion appears useful for future research; it could not be applied easily to a reanalysis of the existing research because many of the investigators have not presented the complete IA matricies or i/d ratios by class in their reports.

Generic Skills of Teaching. Despite the acceptance of individual differences in education, process-product studies have still been designed as if there were one set of effective behaviors that could be applied to all students. One alternative approach is to use analysis of variance in which teachers are classified as high, middle, and low on a number of behaviors, and the class mean achievement scores are used as the cell entries. Another analytic procedure, proposed by Gage and others (personal communication), is to develop a scoring scheme for a heirarchy of teacher behaviors. For example, a heirarchy might be developed in which the relevance of the instruction to the criterion test is considered first, then the cognitive level of the interaction, and then the level of affective interactions. In such a situation, high positive affective behaviors by the teacher might not influence student cognitive growth if the first two conditions were not met, and therefore the scoring scheme would give less weight to teacher affective behaviors.

Almost all the process-product studies have focused upon the relationship of teacher behavior to the class mean. Few investigators have focused on the "personality" or "learning style" of subgroups of learners, or have stratified classes according to the initial knowledge of aptitude of the students. [For a discussion of analyses of main effects and interaction effects, see Walberg (1970b).] For an example of the study of subgroups within a class see Anderson (1970).

There is also the possibility that certain teaching behaviors have differential effectiveness for different types of materials and for students of different ages. Unfortunately, there are not enough studies in any subject area even to begin to suggest different patterns of effectiveness for different materials and grade levels. Finally, we must remain aware of the possibility that teaching and learning are so idiosyncratic that we shall never find anything approaching a set or sets of effective procedures.

Clearer interpretation and review of current and future process-product studies will be possible if investigators include in their reports class means, standard deviations, and the major raw data (such as IA matrices and class means residual gain scores). Such a suggestion has already been made by Thomas Evans (University of Oregon) and has been a requirement Edmund Amidon has imposed on his graduate students. If the inclusion of all such data becomes unwieldy in a document intended for limited distribution, then the data could be deposited in a document center. If such additional information becomes available then the types of reanalyses described above, and also the ones described in the section on analysis of data, will be possible.

Because of the incompleteness of data reported and the questionable reliability of coders across investigations, current research employing systematic observation of classroom behavior might be characterized as a shift from high-inference to medium-inference, a shift from subjective to relatively objective observation. The next shift should be toward greater precision in recording, reporting, and analyzing results.

Experimental studies in Interaction Analysis. Because of the large number of studies in which Interaction Analysis has been used as the observation instrument, and because of the popularity which this observational system has had as a teacher training instrument, there is a need for experimental studies using IA. Unfortunately, many of the existing experimental studies (e.g., Amindon and Flanders, 1961; Schantz, 1963) have limited external validity in that only one teacher, the experimenter, role-played both the indirect and the indirect teacher. In addition, the level of indirectness (e.g., percent of tallies in Category 3) in the indirect condition, and the level of directness in the direct condition were far greater than the levels which occurred in normally indirect classrooms (see matrix in Flanders, 1965).

The critical experimental study in IA would be to select teachers who had low i/d ratios and whose classes were low in residual achievement gain the previous year. Half of these teachers could be trained to increase their i/d ratio, and the effect of such training upon student achievement (compared to the control group) could be assessed. Other experimental studies in this area could focus on optimal indirectness and directness. Teachers could be trained to exhibit i/d ratios above 2, ratios between 1 and 2, and ratios below 1, and the differential effectiveness of the behavior patterns could be assessed. Until studies such as these are completed, the usefulness of attempting to change a teacher's behavior so that his i/d ratio is higher is questionable.

Experimental studies also appear to be the only way to determine the effects of teaching behaviors whose frequency of occurrence is relatively rare, but which are considered important for student achievement. Such rare behaviors include using student ideas by expanding upon what a student said, and asking questions which require analysis, synthesis, or evaluation.

## Reporting Results

In the proposal for this grant were statements that I would try to "delineate optimal groupings of teacher behaviors for different types of outcomes" and identify promising skills in clear unambiguous terms, providing coding instructions which can be used to identify the frequency of use of these skills in a training of classroom situation. Such plans now appear to have been unrealistic.

One problem which currently precludes drawing any conclusions about optimal frequencies of behaviors for certain outcomes, or presenting promising skills in unambiguous terms is the lack of sufficiently clear coding instructions in most of the reports. The problem of clear coding instructions was discussed in Chapter One under "Inter-investigation reliability." The major point made there was that because of unstated ground rules, different investigators using the same observational category system might obtain different results. Without clear descriptions of the ground rules, comparisons among studies are hazardous. Comparison will be facilitated if future process-product studies contain more specific descriptions of the behaviors that are included within any category. Excellent examples of this type of specificiation were contained in the final reports by Spaulding (1965), Snider (1966), and Bellack et al. (1966).

## Experimental Classroom Studies

The results of process-product studies must be treated with caution because they are correlational, not experimental studies. The results of such studies can be deceptive in that they suggest causation although the teacher behaviors which are related to student achievement may be only minor indicators of a complex of behaviors that we have not yet identified. Although hypotheses derived from process-product studies have some usefulness in teacher training programs, experimental studies are the only clear procedure for validating these hypotheses. Researchers and educators in teacher education appear to be unaware of the tentativeness and limitations of the results of process-product studies.

In order to develop acceptable conclusions on whether any of the significant variables identified in Chapters Two through Five should be taught to teacher-trainees or in-service teachers, experimental studies are needed in which teachers are trained to exhibit these behaviors, and the effect of such training on student achievement is assessed. Some of the elements in the design of such studies include: (a) the teacher (or classroom) as the statistical unit of analysis, (b) random assignment of teachers and students to treatment (s), (c) collection of observational data on the behavior of teachers in the experimental, comparison, and/or control classrooms, and (d) assessment of student performance on a variety of end-of-course tests. The comparison and/or control teachers would either follow their normal teaching procedures or provide a specified, alternative instructional procedure.

Such studies are rare. To date, I have found no more than 10 studies which satisfy the above criteria. The scarcity of such studies is not surprising because conducting them involves all the problems of conducting process-product studies, plus additional problems of administration and teacher training.

In this review no attempt was made to synthesize the results of these experimental classroom studies because they are so rare and so varied in the treatment variable. The teacher behaviors which have been studied include asking questions on a higher cognitive level (Rogers and Davis, 1970), using more praise and support of student ideas (Carline, 1970), and teaching a mathematics unit in a discovery or expository manner (Worthen, 1968a,b). The results of these studies were cited when the process-product studies which employed seemingly similar variables were discussed in Chapters Two through Five, and additional results are cited in the Appendix, but there are too few studies to merit a separate review, at present. However, the reader should be aware that a number of seemingly important variables which were identified through correlational studies may not replicate in experimental studies.

## Conclusion

In comparison with the energy and money expanded on the training of teachers, or the development and promotion of educational innovations, on the development of instructional materials, and on the work in laboratory studies of human learning, there have been few well designed correlational or experimental studies of classroom instruction. Reports on laboratory research on meaningful human learning of school subjects are usually concluded with a few paragraphs on "implications for teaching," but these implications are seldom implemented in a teacher training program, much less studies in a systematic fashion when teachers are the mediators of instruction. Most studies on classroom instruction have been conducted by doctoral candidates, and there have been only a few large-scale experimental or correlational studies on teacher behavior and student achievement. Because of this lack of research, we have little knowledge of the relationship between teacher behavior and student growth. Given the number of excellent investigators in the field of education and the amount of research being conducted in natural settings, such a lack of reported studies is shameful. Perhaps this review will help more investigators to become involved in this research.

There have been too few studies of teacher behavior related to student achievement to permit any conclusions on the validity of this type of research. Perhaps when 60 to 80 studies have been completed by investigators using some of the more promising suggestions in this review, we can consider the usefulness of these studies more closely. But future results may not be any clearer than those we have so far. First, we may continue to have trouble identifying the behaviors of good teachers because they are idiosyncratic. A wide range of superior teaching behaviors may be distributed among the superior teachers so that no single behavior or group of behaviors emerges either as a correlate of good teaching or as a discriminating variable.

Second, too many potentially influential variables are not being considered in studies employing systematic observation. These variables include the textbooks and supplementary materials, the organization of the lesson and sequencing of the materials, the cognitive learning style of the individual pupils, and the influence of the entire school environment upon academic achievement. It may be unrealistic to expect that the results of future studies employing systematic observation will be any stronger than the present ones. The author of a major study in this review, has recently written in personal correspondence:

> We keep thinking that any time now we ought to be over the hill and things ought to be easing off, but it never seems to happen; the hill seems to be getting steeper.

# Appendix

This appendix contains a description of the 11 variables which appear most promising for future research on teacher behavior and student achievement, together with a summary of the results obtained when these variables were studied. These variables have been studied through the use of classroom observational category systems, and/or observational rating scales which were used to rate the teacher's classroom behavior on more general (High inference) behaviors. All of the results obtained using observational category systems have already been described in detail in the preceding chapters. This appendix is a summary of the variables this reviewer regarded as most promising given the data in the preceding chapters. A number of variables (such as "clarity") have been primarily studied through the use of observational rating scales and therefore have not been covered in this review. The reader might be interested in learning of the promising variables which have been studied through rating scales (or high inference procedures), and therefore these results are also presented in this appendix.[1] Whenever possible, experimental classroom studies relevant to these variables are also cited. These experimental studies are primarily limited to those in which a number of classrooms received the experimental and the control or comparison treatment. Studies in which one classroom was compared with another classroom are not included.

The 11 variables are:

1. Clarity
2. Variability
3. Enthusiasm
4. Task orientation
5. Student opportunity to learn criterion material
6. Use of student ideas and general indirectness
7. Criticism
8. Use of structuring comments
9. Types of questions
10. Probing
11. Level of difficulty of instruction

---

[1] The research on high inference variables was funded through a grant to the reviewer from the International Association for the Evaluation of Educational Achievement (IEA). This appendix is taken, without revision, from a chapter which the reviewer (and Norma Furst of Temple University) prepared for a book on teacher education which was edited by B.O. Smith and which will be published by Prentice-Hall and the American Educational Research Association.

The strongest results were obtained on the first five variables; the results were less conclusive on the last six variables.

In the summary below, whenever the term "counting" is used, the referant is to studies reviewed in this report in which observational category systems were used to code teacher behavior. The term "rating" is always used to refer to studies in which rating scales were used (such as a rating of the amount of "clarity" a teacher has shown in his lessons).

## 1. Clarity[1]

The cognitive clarity of a teacher's presentation has been studied in seven investigations in which student or observer rating were used. The investigators used different descriptions of clarity:

a) "clarity of presentation" (Belgard et al., 1968; Fortune, 1967; Fortune et al., 1966).

b) whether "the points the teacher made were clear and easy to understand" (Solomon et al., 1963).

c) whether "the teacher was able to explain concepts clearly...had facility with her material and enough background to answer her children's questions intelligently" (Wallen, 1st grade, 1966; Wallen, 3rd grade, 1966).

d) whether the cognitive level of the teacher's lesson appeared to be "just right most of the time" (Chall and Feldman, 1966).

---

[1] The reader should note that all of the studies cited below employed a number of variables as dependent measures, and the results of these studies appear in more than one place. For example, one study of first grade instruction (Wallen, 1st grade, 1966) appears below under the review of "clarity," and also under "task orientation," because both variables were significant in that study. The studies are identified by the name of the investigator, and a reference such as "Fortune, 1967," or "Wallen, 1st grade, 1966" refers to the identical study whenever the same reference is used.

Significant results on at least one criterion measure were obtained in all seven studies. In those studies for which simple correlations were available, the significant correlations ranged from .37 to .71.

Unfortunately, we are uncertain as to the low-inference behaviors which comprise clarity. In studies employing low-inference behaviors, investigators found that the most effective teachers (a) spent less time answering student questions which require interpretation of what the teacher said (Solomon et al., 1963), (b) phrased questions so that they were answered the first time without additional information or additional questions interspersed before the student responded (Wright and Nuthall, 1970), and (c) used fewer "vagueness words" such as "some," "many," "of course," and "a little." (Hiller et al., 1969). Future research might be directed at determining those low-inference behaviors whose frequency of occurrence correlates with ratings on clarity. Once these behaviors are identified, they can be taught in a teacher education program, and the effects of teacher use of the behaviors on student achievement can be assessed.

Another high-inference variable, namely organization, may be similar to clarity because in the study by Solomon et al. (1963) student and observer ratings on "clarity of the lesson," "coherence of the lesson" and "organization of the lesson" all loaded on the same significant factor. The organization of the lesson has also been studied using observer or student ratings on the item "organization of the lesson" (Belgard et al., 1968; Fortune, 1967; Fortune et al., 1966), and student ratings on seven items scales which included items such as, "There is a great deal of confusion during class meetings" (Anderson and Walberg, 1968; Walberg and Anderson, 1968; Walberg, 1969).

Positive relationships between ratings on the behavior labeled "organization: and regression-adjusted student achievement scores were obtained in all the studies mentioned above. Significant correlations between ratings on organization and at least one student achievement measure were obtained in four of six independent studies (Anderson and Walberg, 1968; Belgard et al., 1968; Fortune, 1967; Solomon et al., 1963). The significant correlations ranged from .34 to .67.

Future research will be necessary to determine the specific behaviors which comprise "clarity" or the training procedures which are most likely to achieve high ratings on the clarity of their presentation.

## 2. Variability

A number of studies focused on the teacher's use of variety or variability during the lesson. One investigator (Anthony, 1968) counted the variety of instructional materials, types of tests, and types of teaching devices used by the teacher. Another investigator (Lea, 1964) asked teachers to mark daily checklists on the number of different activities and materials used during social studies lessons. In two studies the investigators coded the cognitive level of classroom discourse and expressed these frequency counts as ratios so that the teacher who employed more cognitive variation in the discourse received higher scores (Furst, 1967; Thompson and Bowers, 1968). Significant results favoring variability were obtained on at least one criterion measure in all four studies.

Other investigators asked students or observers to mark rating scales on (a) the teacher's flexibility in procedure (Solomon et al., 1963), (b) whether the teacher was "adaptable" or "inflexible" (Fortune, 1967), and (c) the amount of extra equipment, books, displays, resource materials, and student activities (Torrance and Parent, 1966; Walburg, 1969). Significant results relating flexibility or abundance were obtained in all four studies. In the studies for which simple correlations were available, the correlations ranged from .24 to .54.

Both high-inference and low-inference correlational studies have indicated that student achievement is positively related to classrooms where a variety of instructional procedures and materials are provided, and where the teacher varies the cognitive level of discourse and of student tasks. It seems worthwhile to study experimentally the effects of training teachers to use this variety.

A variable such as variety appears to be distinct from "flexibility" as defined in recent studies. Flexibility has been studied by counting any form of variation in teacher behavior. For example, Soar (1966) defined flexibility as the number of cells in an Interaction Analysis matrix necessary to account for 60 percent of the tallies. A teacher who used a large number of different cells in the 100 cell matrix would have a high flexibility score. Of eight studies of flexibility, none yielded significant results (Flanders, 2nd grade, 1970; Flanders, 4th grade, 1970; Flanders, 6th grade, 1970; Flanders, 7th grade, 1970; Flanders, 8th grade, 1970; Snider, 1966; Soar, 1966; Vorreyer, 1965). In contrast, in studies of variability, not just any change was counted, but, rather, changes of particular kinds were noted.

## 3. Enthusiasm

Teacher enthusiasm has been assessed by

a) observer ratings on paired adjectives such as "stimulating vs dull" or "original vs stereotyped," or observer ratings on the extent to which the teacher was "interesting and/or dynamic" (Fortune, 1967; Kleinman, 1964; Wallen, 1966)

b) observer estimation of the amount of vigor and power exhibited by the teacher during classroom presentation (Solomon et al., 1963).

c) student ratings on the teacher's involvement, excitement, or interest regarding his subject matter (Solomon et al., 1963).

Significant results relating enthusiasm to studnet achievement on at least one criterion measure were obtained in all five studies in which the variable was studied (rs= .36 to .62), and all non-significant results were in a positive direction (rs= .10 to .30) (Fortune, 1967; Kleinman, 1964; Solomon et al., 1963; Wallen, 1st grade, 1966; Wallen 3rd grade, 1966).

Although the specific, low-inference behaviors which comprise enthusiasm have not yet been identified, the results from correlational and experimental studies suggest that movement, gesture, and voice inflections comprise at least part of this variable (see Rosenshine, 1970d). There is also a hint that mixtures of teacher questions, especially the use of questions calling for interpretation of facts, may be part of the constellation perceived as enthusiasm. New studies should be conducted to determine the low-inferance behaviors which comprise enthusiasm.

It may, however, be possible to train teachers to be more enthusiastic even if we do not know the low inference behaviors. In an experimental study (Mastin, 1963) 20 teachers were given identical materials and told to teach one lesson with enthusiasm and the other without enthusiasm. According to the report, the teachers did not receive further training. The student scores on posttest following these lessons consistently and significantly favored the lessons taught with enthusiasm. Unfortunately, there was not observation of the teachers' classroom behavior.

## 4. Task oriented and/or businesslike

In seven investigations, rating scales were used to estimate the degree to which a teacher was task oriented, achievement oriented, and/or businesslike. Unfortunately, the combination of these studies under one label is hazardous because there is no way to determine whether the different rating scales used can be combined under one category labeled "task oriented and/or businesslike."

In two studies the investigators asked observers to rate the teachers using the paired adjectives which Ryans (1960) identified as comprising "Pattern Y: Businesslike", evading-responsible, erratic-steady, disorganized-systematic, excitable-poised (Fortune, 1967; Kleinmen, 1964). In another study (Chall and Feldman, 1966) the teachers of high achieving classes were rated by observers as emphasizing the stimulation of thought rather than information and skills. In two studies (Wallen, 1st grade, 1966; Wallen, 3rd grade, 1966) "achievement oriented teachers" were rated as being concerned that students learn something, in contrast to students enjoying themselves. In the sixth study students rated their teacher on the extent to which the teacher encouraged the class to work hard and to do independent and creative work (Torrance and Parent, 1966).

Significant results on at least one criterion measure were obtained in all six of the above studies (rs = .42 to .61). In the single study which yielded non-significant results (Beiderman, 1964), student ratings on "task oriented" behavior were not analyzed separately but were combined with student ratings on the teacher's "teacher centered" or "pupil centered" behavior.

Ratings on task orientation may be a significant correlate of student achievement because "you get what you teach for." That is, those teachers who focused upon the learning of cognitive tasks obtained the highest student achievement in this area; those teachers who focused on other activities in the hope that cognitive growth would be obtained indirectly, were less successful. The above extrapolation could be studied by using category systems to determine whether the teachers who are rated high in task-orientation also spend more class time on cognitive tasks and use more cognitive reinforcers with their students.

It may be possible to train teachers to be more task oriented without knowing the low-inference behaviors which comprise this variable. In one experiment (Wittrock, 1962) one group of student teachers was told that their grade in an educational psychology course would be based upon the gain their students attained in American History as compared to the gain attained by classroom students of the control student teachers. The students of the experimental teachers achieved significantly superior growth on a standardized achievement test to that of the students of the control teachers. Unfortunately, no observations were made of the classroom behaviors of the teachers in this experiment.

## 5. Student Opportunity to Learn Criterion Material

A major question in research of this type is whether the criterion instrument was relevant to the instruction. When the students are given a standardized pretest and posttest on reading, and the behaviors of the teacher are correlated with adjusted gain scores, the investigators seldom know whether the material on the posttest was indeed covered in the lessons.

In three investigations an attempt was made to assess the relationship between the material covered in the class and the class criterion score. Two investigators (Rosenshine, 1968; Shutes, 1969) inspected typescripts of fifteen minute lessons to determine the extent to which the material required to answer the posttest was covered in the lesson. A third investigator related the amount of time spent on various topics within four hour-long lessons to student achievement on these topics (Bellack, 1966). In a cross-cultural study involving over 300,000 students in twelve countries, the teachers were shown the criterion test and were asked to rate whether "all or most (at least 75%)," "some (25% to 75%)", or "few (less than 25%)" of their students had the opportunity to learn the type of problem exemplified by each test item (Husen, 1967).

Significant correlations between "opportunity to learn" and student achievement were obtained in three of the four studies (Husen, 1967; Rosenshine, 1968; Shutes, 1969) (rs = .16 to .40). The significant correlations in the cross-cultural study were obtained for each of four groups of students and represent the median within country correlation (Husen, 1967). That significant results did not occur in the fourth study (Bellack et al., 1966) may have been because the test items themselves were not studied.

Overall the correlations between measures of opportunity to learn and student achievement are positive, significant, and consistent. Note that in the largest of these studies (Husen, 1967) the teachers had never seen the test material before and were asked whether students had had an opportunity to learn material of this type. These results suggest that there is a positive correlation between the types of cognitive processes the students had an opportunity to learn and student performance on the international mathematics test. (However, the correlations are based on teacher reports and must be corroborated by direct observation.) One implication for teacher education is that it is important to orient teachers towards cognitive classroom activities if we wish to enhance student cognitive growth. Experimental studies that test these ideas would be desirable.

The high, significant correlations obtained in two other studies discussed above (Rosenshine, 1968; Shutes, 1969) can be interpreted as measuring the degree to which teachers trained their students on the criterion items. Such results have implications for the statistical analyses of studies of teaching and will be discussed in the next section.

## 6. Use of Student Ideas and General Indirectness

The behavior, "teacher use of student ideas," was originally developed by Flanders (1965) and appears as Category 3 of his Interaction Analysis (IA) system. Although considerable correlational and descriptive research has been conducted using IA, the variable "use of student ideas" remains ambiguous. Flanders (1970) has attempted to solve the problems of definition by dividing this category into five sub-categories of behaviors:

1. Acknowledging the student's idea by repeating the nouns and logical connectives he has expressed.
2. Modifying the idea by rephrasing it or conceptualizing it in the teacher's own words.
3. Applying the idea by using it to reach an inference or take the next step in a logical analysis of a problem.
4. Comparing the idea by drawing a relationship between it and ideas expressed earlier by the students or the teacher.
5. Summarizing what was said by an individual student or group of students.

Flanders reported (personal communication) that at least 60 per cent of the behaviors classified as Category 3 consist of simple repetition by the teacher of what the student said.

Eight studies have been found in which counts of total use of student ideas and/or counts of extended (more than three seconds) use of student ideas were correlated with measures of student achievement. A significant bivariate correlation between teacher use of student ideas and student achievement was not obtained in any study. However, in 7 of the 8 studies correlations were positive (Flanders, 4th grade, 1970; Flanders, 6th grade, 1970; Flanders, 7th grade, 1970; Flanders, 8th grade, 1970; Perkins, 1965; Soar, 1966; Wright and Nuthall, 1970) (rs= .17 to 40). The consistency of these results suggest that the variable, teacher use of student ideas, appears important enough to warrant more intensive study.

Another variable derived from the Flanders' Interaction Analysis matrix has been labeled "indirectness." It consists of the combined frequencies of teacher behaviors labeled (a) acceptance of student feeling, (b) praise or encouragement, and (c) use of student ideas. Such behaviors may be similar to the variable labeled "emotional climate" (Medley and Mitzel, 1959).

The results of six studies utilizing this variable were similar to those obtained when "teacher use of student ideas" was studied. Significant results were obtained in one study (Flanders, 6th grade, 1970; Flanders, 8th grade, 1970; Medley and Mitzel, 1959) (rs = .12 to .41). Because the variable "teacher use of student ideas" is part of the more general variable "indirectness," both variables appear to be useful for future research.

A third variable, the ratio of "indirect" to "direct" behaviors, also appears to be useful for future study. This ratio has been significantly related to student achievement in only one study (LaShier, 1967) but positive correlations were obtained in 11 of 13 investigations (rs = .12 to .41).

There have been four experimental classroom studies in which teachers were trained to be more supportive, their classroom behaviors were observed, and class achievement scores were compared with those obtained in classrooms which received a contrast or control treatment. Unfortunately, the results were not statistically significant, nor was there a discernable trend in the four studies (Carline, 1969; Gunnison; 1968; Herman et al., 1969; Miller, 1966).

## 7. Criticism

Teacher use of behaviors labeled "criticism" or "control" has been one of the most frequently counted variables in process-product research. Seventeen studies were reported in which observers counted these behaviors. Many of the investigators used more than one measure of criticism. For example, in five separate studies one investigator computed counts of (a) total teacher use of criticism and giving of directions, (b) extended (more than three seconds in duration) teacher criticism and giving of directions, and (c) teacher criticism or directions in response to student comments (Flanders, 1970). Another investigator (Hunter, 1968) developed separate categories for hostile or strong disapproval, neutral or mild disapproval, directive statements related to school, and teacher justification of authority. Other investigators (Harris and Serwer, 1966; Harris et al., 1968) divided teacher criticism into negative motivation and control.

Significant negative relationships between some form of criticism and at least one criterion measure were obtained in 6 to 17 studies (Anthony, 1967; Flanders, 7th grade, 1970; Harris et al., 1968; Hunter, 1968; Soar, 1966; Wallen, 1st grade, 1966) (rs = -.38 to -.61). Both positive and negative relationships were obtained in two studies which employed factor analysis (Perkins, 1965; Spaulding, 1965), and significantly positive results were obtained in one study (Harris and Serwer, 1966) (rs = .28 to .29). On the whole, there is a trend for significant negative relationships between teacher criticism and student achievement, but the results are not as strong as some of the other variables discussed in this paper.

If only the direction of the correlation is considered, negative correlations between all observed measures of criticism and all measures of achievement were obtained in 12 of the 17 studies (Anthony, 1967; Cook, 1967; Flanders, 4th grade, 1970; Flanders, 6th grade, 1970; Flanders, 7th grade, 1970; Flanders, 8th grade, 1970; Harris et al., 1968; Hunter, 1968; Soar, 1966; Wallen, 1st grade, 1966; Wallen, 3rd grade, 1966; Wright and Nuthall, 1970). These correlations ranged from -.04 to -.62. Positive correlations between all measures of criticism and all measures of achievement were obtained in two studies (Harris and Serwer, 1966; Morsh, 1956), but these correlations tended to be small (rs from .05 to .29). Both positive and negative relationships between criticism and achievement were obtained in three of the 17 studies (Flanders, 2nd grade, 1970; Perkins, 1965; Spaulding, 1965). In sum, the direction of the correlations shows a strong trend for a negative relationship between criticism and student achievement.

In 16 of the studies, it is possible to compare the relationships of different types or intensities of criticism to student achievement. For example, the results on "mild disapproval" can be compared with those on "strong disapproval" (Hunter, 1968), or the results on "rejecting a student response" can be compared with "teacher criticizes or justifies authority" (Perkins, 1965). In 10 of the 17 studies, the stronger form of criticism had a higher negative correlation with achievement than the milder form. Thus, teachers who use extreme amounts or forms of criticism usually have classes which achieve less in most subject areas.

In no study was there a significant negative correlation between mild forms of criticism or control and student achievement. Such mild forms include telling a student that his answer was incorrect or providing academic directions. Thus there is no evidence to support a claim that teachers should avoid telling a student he was wrong or should avoid giving academic directions.

Variables such as teacher use of differing forms of approval and disapproval are frequently used as performance criteria in teacher education programs. But it is impossible to make any specific recommendations on the implications of this research for teacher training for two reasons. First, in correlational studies such as these we do not know if the teacher's use of criticism is self-initiated, results from the character of the students, or results from an interaction of the teacher and students. Second, we do not know if the variables labeled as approval or disapproval in one study are comparable with those so labeled in another. In future research there is a need to subdivide these variables into smaller units such as increasing levels of affect and to design observational systems that enable us to record the context in which these behaviors occur.

## 8. Use of structuring comments

Investigators who have counted the use of teacher "structuring" statements generally refer to statements designed to provide an overview or a cognitive scoffolding for what is to happen or has happened. Such statements have been identified at the start and at the end of lessons and at the start and end sections of lessons. Teacher statements which precede a question, statements which summarize an interchange, the use of a clear signal to indicate when one part of a lesson ends and another begins, and verbal markers of importance (e.g., "Now get this") are among the diverse procedures used to identify structuring. Teacher structuring statements have been counted in four investigations, and significant results were obtained in all four (Furst, 1967; Penny, 1969; Soar, 1966; Wright and Nuthall, 1970). Structuring statements were also cited in two investigations in which the significance levels were not given (Crossman and Olson, 1969; Fortune, 1967). Although each investigator gave fairly precise operational definitions of the variable, the category systems used were so different that we cannot make comparisons of the results.

In three studies in which raters estimated the adequacy of the beginning or the ending of the lesson, there were a significant correlations (rs = .35 to .69) between ratings for either the beginning or the ending of the lesson, there were significant correlations (rs = .35 to .69) between ratings for either the beginning or the end of the lesson and the criterion measure (Belgard et al., 1968; Fortune, 1967; Fortune et al., 1966). Although all correlations were positive, the correlations were significant for both the beginning and the end of the lesson in only one study (Fortune, 1967). Unfortunately, we are unable to determine whether there is any relationship between the ratings given to the beginning and end of the lesson, and the various counts of structuring.

The results to date indicate that the various forms of structuring merit further study, but it is impossible to synthesize the results in a manner which can be translated into teaching competencies. Only fragmentary hints for teacher training programs can be offered, such as, consider providing a moderate number of statements before asking a question, reviewing at the end of a series of interchanges, using a review at the start or end of a lesson, or providing clear signals as to when one part of a lesson ends and another begins.

## 9. Types of Questions

Two classifications. Several investigators have studied the relationship between teacher use of various types of questions (or varied types of classroom discourse) and student achievement. Most investigators have used a scheme in which questions are classified into two forms. In general, the two forms might be labeled "lower cognitive level," and "higher cognitive level" questions, although few investigators used these specific labels. Each investigator provided fairly clear definitions of his categories, and most investigators tended to classify questions which focused on "what" or "where" as lower level questions and questions of "why" and "how" as higher level questions. However, classifications among investigators overlap in such a way that a question which was classified as lower level in one investigation might have been classified as higher level in another.

Of the seven investigations in which two types of questions were classified, significant results were not obtained in four investigations (Harris and Serwer, 1966; Harris et al., 1968; Perkins, 1965; Wright and Nuthall, 1970). The reports did not present sufficient detail to specify the overall direction of the correlations. Of the three investigations in which significant results were reported, the high achieving teachers asked more "high level" questions in one study (Kleinman, 1964), but asked fewer "open ended" questions in another study (Spaulding, 1965). In the third study, the highest achieving teachers were those who mixed convergent and divergent questions (Thompson and Bowers, 1968).

Thus, the classification of all questions into only two forms has not yielded consistent significant results or any discernable trend.

Multiple Classifications of Discourse. Only two studies were found which used multiple classifications of teacher questions or types of teacher-student discourse. Significant results were obtained in both (Conners and Eisenberg, 1966; Solomon et al., 1963). The studies are not easily compared because they differed widely in design, coding procedures, and focus. Not even a tentative conclusion can be drawn on the relationship between various cognitive levels of discourse and student achievement. The most useful conclusion at this point is that classification of questions and/or types of discourse into three or more types appears to offer greater potential for future research than the use of only two classifications.

## 10. Probing.

The variable "probing" generally refers to teacher responses to student answers in which the teacher responses encourage the student (or another student) to elaborate upon his answer. In one investigation the teacher "elicited clarification in a non-threatening way" (Spaulding, 1965) and in another (Soar, 1966) teachers were scored as encouraging "interpretation, generalization, and solution" if they asked such a question, or if they responded to a student in such a manner. In a third investigation (Wright and Nuthall, 1970) various types of teacher responses were counted, such as redirection of the question to another student, or the asking of another question to the student who first answered. Significant results were obtained in all three studies (rs = .29 to .54), but the variety of methods used to record such behavior precluded any synthesis of the results. We can conclude only that further study of such teacher behaviors appears warranted.

## 11. Level of Difficulty of Instruction

Student perceptions of the difficulty of the instruction have been assessed in four studies through student questionnaires. One investigator (Walberg, 1969) used a seven item scale which contained items such as, "The class is best suited for the smartest students." However, two of the items in the difficulty scale may refer to the aptitude or brightness of the students in the class: "Students in the class tend to be much brighter than those in the rest of the school;" "Many students in the school would have difficulty doing the advanced work of the class." Because the challenge of the course and the brightness of the students are both in the same scale, it is impossible to determine from the data whether the measured student perception of "difficulty" is a function of the teacher's approach, the ability of the class, or an interaction of the two.

In the international study cited above, (Husen, 1967) students were asked to rate the difficulty of learning mathematics on a five point scale. In the third study (Nikoloff, 1966) a specially prepared questionnaire was developed to assess how strict the teacher was in demanding high standards in English composition. In the fourth study (Torrance and Parent, 1966) one of the questionnaire items was: This class is one of the hardest in the school.

There was a clear, significant relationship between student perception of difficulty and student achievement in two of the four studies (Torrance and Parent, 1966; Walberg, 1969) (r = .44), and no discernable trend in the other two studies (Husen, 1967; Nikoloff, 1965).

Student perception of level of difficulty appears to be a fascinating area for future study because in two studies perceptions of difficulty were positively related to achievement. However the issue is more complex because in the study with the strongest results (Walberg, 1969) mean perceptions of difficulty in this special physics program were lower than perceptions of difficulty of the regular physics program.

## Summary of process-product results

Summary of strongest findings. Of all the variables which have been investigated in process-product studies to date, five variables have strong support from correlational studies, six variables have less support, but appear to deserve future study. The five variables which yielded the strongest relationships with measures of student achievement are: clarity, variability, enthusiasm, task orientation and/or bussiness-like behavior, and student opportunity to learn. The six less strong variables are: use of student ideas and/or teacher indirectness, use of criticism, use of structuring comments, use of multiple levels of discourse, probing, and perceived difficulty of the course. The relationships are positive for ten of the variables and negative for use of criticism.

Summary of non-significant results. At first glance, the above list of the strongest findings may appear to represent mere educational platitudes. Their value can be appreciated, however, only when they are compared to the behavioral characteristics, equally virtuous and "obvious" which have not shown significant or consistent relationships with achievement to date. These variables, which are taken from the larger reviews (Rosenshine 1970a,b), are listed below, and the method by which they were assessed follows in parenthesis: non-verbal approval (counting), praise (counting), warmth (rating), ratio of all indirect behaviors to all direct teacher behaviors, or the I/D ratio (counting), flexibility (counting), questions or interchanges classified into two types (counting), teacher talk (counting), student talk (counting), student participation (rating), number of teacher-student interactions (counting), student absence, teacher absence, teacher time spent on class participation (rating), teacher experience, and teacher knowledge of subject area. It is possible that future studies employing improved designs and improved analyses of the data, or future reviews of the same literature may yield somewhat different conclusions. However, such caution works both ways--one cannot claim that the above non-significant variables are correlates of student achievement until he can marshal supportive data.

## References

Ackerman, W. I. Teacher competence and pupil change. Harvard Educational Review, 1954, 24, 213-289.

Amidon, E. J., & Flanders, N. A. The effects of direct and indirect teacher influence on dependent-prone students learning geometry. Journal of Educational Psychology, 1961, 52, 286-291.

Amidon, E. J., & Flanders, N. A. The role of the teacher in the classroom. Minneapolis: Association for Productive Teaching, 1967.

Amidon, E. J., & Simon, A. Teacher-pupil interaction. Review of Educational Research, 1965, 25, 130-140.

Amidon, E. J., Amidon, P., & Rosenshine, B. Skills development in Teaching Work Manual. Minneapolis: Association for Productive Teaching, 1969.

Anderson, G. J. Effects of classroom social climate on individual learning. American Educational Research Journal, 1970, 7, 135-153.

Anderson, G. J., Walberg, H. J., & Welch, W. W. Curriculum effects on the social climate of learning: A new representation of discriminant functions. American Educational Research Journal, 1969, 6, 315-329.

Anderson, G. J., & Walberg, H. J. Classroom climate and group learning. International Journal of Educational Sciences, 1968, 2, 175-180. (ERIC: ED 015 156.)

Anderson, H. H. The measurement of domination and of socially integrative behavior in teachers' contacts with children. Child Development, 1939, 10, 73-89.

Anderson, R. C. Control of student mediating processes during verbal learning and instruction. Review of Educational Research, 1970, 3, 349-371.

Anderson, R. C., et al. (Eds.). Current research on instruction. Englewood Cliffs, N. J.: Prentice-Hall, 1969.

Anthony, B. C. M. The identification and measurement of classroom environmental process variables related to academic achievement. Unpublished doctoral dissertation, Univ. of Chicago, 1967.

Aspy, D. N. *The effect of teacher inferred self concept upon student achievement*. Paper presented at the meeting of the American Educational Research Association, Feb. 1969. Gainesville: Univ. of Florida, Dept. of Education. (Mimeo)

Ausubel, D. P. *The psychology of meaningful verbal learning: An introduction to school learning*. New York: Grune and Stratton, 1963.

Beiderman, D. D. Relationship between teaching style and pupil behavior. Unpublished doctoral dissertation, Univ. of Calif. at Los Angeles, 1964. University Microfilms No. 64-12, 180.

Belgard, M., Rosenshine, B., & Gage, N. L. The teacher's effectiveness in explaining: Evidence on its generality and correlation with pupils' ratings and attention scores. In N. L. Gage, et al., *Explorations of the teacher's effectiveness in explaining*. Technical Report No. 4, Stanford, Calif.: Stanford Center for Research and Development in Teaching, Stanford Univ., 1968. (ERIC: ED 028 147.)

Bellack, A. A., Kleibard, H. M., Hyman, R. T., & Smith, F. L. *The language of the classroom*. New York: Columbia Univ. Press, 1966.

Birkin, T. A. Toward a model of instructional processes. In I. Westbury and A. A. Bellack (Eds.), *Research into classroom processes*. New York: Teachers College Press, in press.

Bond, G. L., & Dykstra, R. The cooperative research program in first-grade reading instruction. *Reading Research Quarterly*, Summer, 1967, 2, 1-142.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs in research on teaching. In N. L. Gage (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 171-247.

Carline, J. L. An investigation of the relationship between various verbal strategies of teaching behavior and achievement of elementary school children. Unpublished doctoral dissertation, Syracuse Univ., Syracuse, New York, 1969.

3

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-732.

Chall, J. S., & Feldmann, S. C. A study in depth of first grade reading. Cooperative Research Project No. 2728. New York: The City College of the City Univ. of New York, 1966.

Christensen, C. M. Relationships between pupil achievement, pupil affect-need, teacher warmth, and teacher permissiveness. Journal of Educational Psychology. 1960, 59, 169-174.

Coats, W. D. Investigation and simulation of the relationships among selected classroom variables. U. S. Office of Education Research Project No. 6-8330. Ann Arbor, Mich.: Univ. of Mich., 1966. Also available as an unpublished doctoral dissertaition, Univ. of Mich., 1966.

Coats, W. D., & Smidchens, U. Audience recall as a function of speaker dynamism. Journal of Educational Psychology, 1966, 57, 189-191.

Conners, C. K., & Eisenberg, L. The effect of teacher behavior on verbal intelligence in operation headstart children. U. S. Office of Economic Opportunity Headstart Contract No. 510. Baltimore: Johns Hopkins Univ. School of Medicine, 1966. (ERIC: ED 010 782.)

Cook, R. E. The effect of teacher methodology upon certain achievements of students in secondary school biology. Unpublished doctoral dissertation, Univ. of Iowa, Iowa City, 1967.

Cronbach, L. J., & Furby, L. How should we measure "change"--or or should we? Psychological Bulletin, 1970, 74, 68-80.

Crossan, D., & Olson, D. R. Encoding ability in teacher-student communication games. Paper presented at the meeting of the American Educational Research Association, Feb. 1969. Toronto: Ontario Institute for Studies in Education.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-402.

Fey, J. Classroom teaching of mathematics. *Review of Educational Research*, 1969, 39, 535-551.

Flanders, N. A. *Teacher influence, pupil attitudes, and achievement.* Cooperative Researach Monograph No. 12, OE-25040. Washington, D. C.: U. S. Government Printing Office, 1965.

Flanders, N. A. *Analyzing teaching behavior*. New York: Addison-Wesley, 1970.

Flanders, N. A., & Simon, A. Teacher effectiveness. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.), New York: Macmillan, 1969.

Fortune, J. C. *The generality of presenting behaviors in teaching preschool chilcren*. Memphis, Tenn.: Memphis State Univ., 1966.

Fortune, J. C. *A study of the generality of presenting behaviors in teaching preschool children*. U. S. Office of Education Project No. 6-8468. Memphis, Tenn.: Memphis State Univ., 1967.

Fortune, J. C., Gage, N. L., & Shutes, R. E. *The generality of the ability to explain*. Paper presented at the meeting of the American Educational Research Association, Feb. 1966. Amherst: Univ. of Mass., College of Education.

Fowler, B. D. Relations of teacher personality characteristics and attitudes to teacher-pupil rapport and emotional climate in the elementary classroom. Unpublished doctoral dissertation, Univ. of S. Carolina, Columbia, 1962.

Furst, N. F. *The multiple languages of the classroom.* Paper presented at the meeting of the American Educational Research Association, Feb. 1967. Also available as an unpublished doctoral dissertation, Temple Univ., Philadelphia, Pa., 1967.

Furst, N. F., & Amidon, E. J. Teacher-pupil interaction patterns in the elementary school. In E. J. Amidon & H. Hough, (Eds.), *Interaction analysis: Theory, research, and application.* Reading, Mass.: Addison-Wesley, 1967.

Gage, N. L. Paradigms for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. Pp. 94-142.

Gage, N. L. Research on cognitive aspects of teaching. *Report of the seminar on teaching: The way teaching is*. Washington, D. C.: Association for Supervision and Curriculum Development and National Educational Association, 1966.

Gage, N. L. Teaching methods. In R. L. Ebel (Ed.), *Encyclopedia of educational research* (4th ed.), New York: Macmillan, 1969.

Gage, N. L., & Unruh, W. R. Theoretical formulations for research on teaching. *Review of Educational Research*, 1967, 37, 358-370.

Gage, N. L., et al. *Explorations of the teacher's effectiveness in explaining*. Technical Report No. 4, Stanford, Calif.: Stanford Center for Research and Development in Teaching, Stanford Univ., 1968. (ERIC: ED 028 147.)

Gallagher, J. J. *Analyses of teacher classroom strategies associated with student cognitive and affective performance*. U. S. Office of Education Project No. 3325. Urbana: Univ. of Ill., 1968. (ERIC: ED 021 808.)

Gallagher, J. J. Three studies of the classroom. In American Educational Research Association Monograph No. 6, *Classroom Observation*. Chicago: Rand-McNally, 1970.

Guggenheim, F. Classroom climate and the learning of mathematics. *Arithmetic Teacher*, 1961, 8, 363-367.

Guilford, J. P. *The nature of human intelligence*. New York: McGraw-Hill, 1966.

Gunnison, J. F. An experiment to determine the effects of changing teacher classroom behavior through training of student-teachers in the use of Flanders interaction analysis system. Unpublished doctoral dissertation, Ariz. State Univ., Tempe, 1968.

Hammachek, D. Characteristics of good teachers and implications for teacher education. *Phi Delta Kappan*, 1969, 6, 341-345.

Harris, A. J., Morrison, C., Serwer, B. L., & Gold, L. *A continuation of the CRAFT project: Comparing reading approaches with disadvantaged urban negro children in primary grades*. U. S. Office of Education Project No. 5-0570-2-12-1. New York: City Univ. of New York, Division of Teacher Education, 1968. (ERIC: ED 010 297.)

Harris, A. J., & Serwer, B. *Comparison of reading approaches in first grade teaching with disadvantaged children (the CRAFT project*. Cooperative Research Project No. 2677. New York: City Univ. of New York, 1966. (ERIC: ED 010 037.)

Harvard Project Physics, 1967 sample. See Anderson & Walberg (1968) and Walberg & Anderson (1968).

Harvard Project Physics, 1968 sample. See Walberg, *American Educational Research Journal* (1969) and Walberg, *Journal of Educational Psychology* (1969).

Heil, L. M., Powell, M., & Feifer, I. *Characteristics of teacher behavior related to the achievement of children in several* elementary grades. Cooperative Research Project No. SAE 7285, 1960. (ERIC: ED 002 843.)

Herman, W. L., Potterfield, J. E., Dayton, C. M., & Amershek, K. G. The relationship of teacher-centered activities and pupil-centered activities to pupil achievement and interest in 18 fifth-grade social studies classes. *American Educational Research Journal*, 1969, 6, 227-240.

Hiller, J. H., Fisher, G. A., & Kaess, W. A computer investigation of verbal characteristics of effective classroom learning. *American Educational Research Journal*, 1969, 6, 661-677.

Hook, E. N. Teacher factors influencing pupil achievement in elementary school English. Unpublished doctoral dissertation, Colo. State College, Greeley, 1965. University Microfilms No. 66-5989.

Hough, J. *Specifications for a comprehensive undergraduate and inservice teacher education program for elementary teachers.* Washington, D. C.: U. S. Office of Education Bureau of Research, 1968. U. S. Government Printing Office PS5.258:58016.

Howe, R. W. The relationship of learning outcomes to selected teacher factors and teaching methods in tenth grade biology classes in Oregon. Unpublished doctoral dissertation, Ore. State Univ., Corvallis, 1964.

Hunter, C. P. Classroom climate and pupil characteristics in special classes for the educationally handicapped. Unpublished doctoral dissertation, Univ. of S. Calif., Los Angeles, 1968.

7

Husen, T. (Ed.). *International study of achievement in mathematics: A comparison of twelve countries, Vol. 2.* New York: John Wiley & Sons, 1967.

Hutchinson, W. L. Creative and productive thinking in the classroom. *Journal of Creative Behavior*, 1967, 1, 419-427.

Jackson, P. *Life in classrooms.* New York: Holt, Rinehart, & Winston, 1968.

Katz, L. G. *Teaching in preschools: Roles and goals.* Document No. 70706-E-AO-U-26. Urbana: National Laboratory on Early Childhood Education, Univ. of Ill., 1969.

Klein, S. Evaluating tests in terms of the information they provide. *Evaluation Comment 2, No. 2.* Los Angeles: Center for the Study of Evaluation, Univ. of Calif., June 1970.

Kleinman, G. General science teachers' questions, pupil and teacher behaviors, and pupils' understanding of science. Unpublished doctoral dissertation, Univ. of Vir., Charlottesville, 1964. University Microfilms No. 65-3961.

Kounin, J. S. An analysis of teachers' managerial techniques. *Psychology in the Schools*, 1967, 4, 221-227.

Lahaderne, H. M. Adaptation to school settings: A study of children's attitudes and classroom behavior. Unpublished doctoral dissertation, Univ. of Chicago, 1967.

LaShier, W. S., Jr. *An analysis of certain aspects of the verbal behavior of student teachers of eighth grade students participating in a BSCS laboratory block.* Paper presented at the meeting of the National Science Teachers Association, April 1966. Also available as an unpublished doctoral dissertation, Univ. of Tex., Austin, 1965.

Lea, H. M. H. A study of some characteristics of the variability of teacher activities in the social studies and pupil response and achievement. Unpublished doctoral dissertation, Univ. of Minn., Minneapolis, 1964. University Microfilms No. 64-9494.

Lewin, K., Lippit, R., & White, R. Patterns of aggressive behavior in experimentally created "social climates." *Journal of Social Psychology*, 1939, 10, 271-299.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: Univ. of Wis. Press, 1967. Pp. 21-38.

Mastin, V. E. Teacher enthusiasm. *Journal of Educational Research*, 1963, 56, 385-386.

McCall, W. A., & Krause, G. R. Measurement of teacher merit for salary purposes. *Journal of Educational Research*, 1959, 53, 73-75.

McDonald, F. J. The influence of learning theories on education: 1900-1950. In E. R. Hilgard (Ed.), *Theories of learning and instruction*. Chicago: Univ. of Chicago Press, 1964. Pp. 1-26.

Medley, D. M., & Mitzel, H. E. A technique for measuring classroom behavior. *Journal of Educational Psychology*, 1958, 49, 86-92.

Medley, D. M., & Mitzel, H. E. Some behavioral correlates of teacher effectiveness. *Journal of Educational Psychology*, 1959, 50, 239-246.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally, 1963. Pp. 247-329.

Meux, M. O. Studies of learning in the school setting. *Review of Educational Research*, 1967, 37, 539-562.

Miller, G. L. An investigation of teaching behavior and pupil thinking. Unpublished doctoral dissertation, Univ. of Utah, Salt Lake City, 1964.

Miller, G. L. Collaborative teaching and pupil thinking. *Journal of Teacher Education*, 1966, 17, 337-358.

Mitzel, H. E. Teacher effectiveness. In C. W. Harris (Ed.), *Encyclopedia of educational research* (3rd ed.), New York: Macmillan, 1960. Pp. 1481-1486.

Morrison, B. M. The reactions of internal and external children to patterns of teaching behavior. Unpublished doctoral dissertation, Univ. of Mich., Ann Arbor, 1966.

Morsh, J. E. *Systematic observation of instructor behavior*. Development Report No. AFPTRC-TN-56-52. San Antonio, Tex.: Air Force Personnel and Training Research Center, Lackland Air Force Base, 1956.

9

Morsh, J. E., Burgess, G. G., & Smith, P. N. Student achievement as a measure of instructor effectiveness. Project No. 7950, Task No. 77243. San Antonio, Tex.: Air Force Personnel and Training Research Center, Lackland Air Force Base, 1955.

Morsh, J. E., & Wilder, E. W. Identifying the effective instructor: A review of the quantitative studies, 1900-1952. Research Bulletin No. AFPTRC-TR-54-44. San Antonio, Tex.: Air Force Personnel and Training Research Center, Lackland Air Force Base, 1954.

Nikoloff, S. E. B. The relationship of teacher standards to the written expression of fifth and sixth grade children. Unpublished doctoral disseration, State Univ. of New York at Buffalo, 1965. University Microfilms No. 66-3520.

Nuthall, G. A. An experimental comparison of alternative strategies for teaching concepts. American Educational Research Journal, 1968, 5, 561-584.

O'Leary, K. D., & Becker, W. C. Behavior modification of an adjustment class: A token reinforcement program. Exceptional Children, 1967, 33, 637-642.

Orme, M. E. J., & Purnell, R. F. Behavior modification and transfer in an out-of-control classroom. Monograph No. 5, Cambridge, Mass.: Center for Research and Development on Educational Differences, Harvard Univ., 1968.

Penny, R. E. Presenting behaviors and teacher success. Unpublished doctoral dissertation, Stanford Univ., Stanford, Calif., 1969.

Perkins, H. V. A procedure for assessing the classroom behavior of students and teachers. American Educational Research Journal, 1964, 50, 240-260.

Perkins, H. V. Classroom behavior and under-achievement. American Educational Research Journal, 1965, 2, 1-12.

Popham, W. J. Performance tests of instructor competence for trade and technical education. U. S. Office of Education Project No. 5-004. 1968. Univ. of Calif. at Los Angeles.

Popham, W. J. Curriculum materials. Review of Educational Research, 1969, 39, 319-339.

Popham, W. J. Validation results: Performance tests of teaching proficiency in vocational education. Paper presented at the meeting of the American Educational Research Association, Feb. 1969. Univ. of Calif. at Los Angeles.

10

Powell, E. R. Teacher behavior and pupil achievement. Paper presented at the meeting of the American Educational Research Association, Feb. 1968. Also available as an unpublished doctoral dissertation, Temple Univ., Philadelphia, Pa., 1968.

Reichenberg-Hackett, W. Practices, attitudes, and valued in nursery group education. Psychological Report, 1962, 10, 151-172.

Rosenshine, B. Objectively measured behavioral predictors of effectiveness in explaining. In N. L. Gage, et al., Explorations of the teacher's effectiveness in explaining. Technical Report No. 4, Stanford, Calif.: Stanford Center for Research and Development in Teaching, Stanford Univ., 1968. (ERIC: ED 028 147.)

Rosenshine, B. To explain: A review of research. Educational Leadership, 1968, 26, 303-309.

Rosenshine, B. Teaching behaviors and student achievement. Stockholm, Sweden: International Association for the Evaluation of Educational Achievement, 1970. (Mimeo.) (a)

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, in press (December). (b)

Rosenshine, B. Evaluation of instruction. Review of Educational Research, 1970, 40, 279-301. (c)

Rosenshine, B. Enthusiastic teaching: Enhanced learning. School Review, 1970, in press (summer). (d)

Rosenshine, B., & Furst, N. F. Current and future research on teacher performance criteria. Paper presented at the meeting of the American Educational Research Association, Mar. 1970. Urbana: Univ. of Ill., College of Education. (Mimeo.)

Rothkopf, E. Z. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. American Educational Research Journal, 1966, 3, 241-249.

Ryans, D. G. Characteristics of teachers. Washington, D. C.: American Council on Education, 1960.

Samph, T. Observer effects on teacher behavior. Paper presented to the meeting of the American Educational Research Association, Feb. 1969. Syracuse, New York: Syracuse Univ.

Schantz, B. M. B. An experimental study comparing the effects of verbal recall by children in direct and indirect teaching methods as a tool of measurement. Unpublished doctoral dissertation, University Park, Pa. State Univ., 1963.

Shannon, J. R. A measure of the validity of attention scores. *Journal of Educational Research*, 1942, 35, 623-631.

Sharp, C. S. A study of certain teacher characteristics and behavior as factors affecting pupil achievement in high school biology. Unpublished doctoral dissertation, Univ. of S. Carolina, Columbia, 1966. University Microfilms No. 66-11, 601.

Shutes, R. E. Verbal behaviors and instructional effectiveness. Unpublished doctoral dissertation, Stanford Univ., Stanford, Calif., 1969.

Simon, A., & Boyer, G. E. (Eds.). *Mirrors for behavior: An anthology of classroom observation instruments*. Philadelphia, Pa.: Research for Better Schools, 1967.

Smith, B. O., Meux, M., et al. *A study of the logic of teaching*. U. S. Office of Education Research Project No. 258. Urbana: Bureau of Educational Research, College of Education, Univ. of Ill., 1962.

Smith, B. O., Meux, M. O., Coombs, J., & Nuthall, G. A. *A tentative report on strategies of teaching*. Urbana: Bureau of Educational Research, College of Education, Univ. of Ill., 1964. (ERIC: ED 029 165.)

Smith, B. O., et al. *A study of the strategies of teaching*. Urbana: Bureau of Educational Research, College of Education, Univ. of Ill., 1967.

Snider, A. M. N. Some relationships between pupil growth in certain basic skills and pupils' perceptions of behaviors of their teachers. Unpublished doctoral dissertation, Univ. of Mich., Ann Arbor, 1965. University Microfilms No. 65-11, 038.

Snider, R. M. *A project to study the nature of effective physics teaching*. U. S. Office of Education Research Project No. S-280. Ithaca, New York: Cornell Univ., 1966. Also available as an unpublished doctoral dissertation, Cornell Univ., 1966.

Soar, R. S. *An integrative approach to classroom learning*. Public Health Service Grant No. 5-R11-MH 01096 and National Institute of Mental Health Grant No. 7-R11-MH 02045. Temple Univ., Philadelphia, Pa., 1966.

Soar, R. S. *Optimum teacher-pupil interaction for pupil growth*. Paper presented at the meeting of the American Educational

Solomon, D. Teacher behavior dimensions, course characteristics, and student evaluations of teachers. American Educational Research Journal, 1966, 3, 35-47.

Solomon, D., Bezdek, W. E., & Rosenberg, L. Teaching styles and learning. Chicago: Center for the Study of Liberal Education for Adults, 1963.

Solomon, D., Bezdek, W. E., & Rosenberg, L. Dimensions of teacher behavior. Journal of Experimental Education, 1964, 33, 23-40.

Solomon, D., Rosenberg, L., & Bezdek, W. E. Teacher behavior and student learning. Journal of Educational Psychology, 1964, 55, 23-40.

Spaulding, R. L. Achievement, creativity, and self-concept correlates of teacher-pupil transactions in elementary schools. Cooperative Research Project No. 1352. Hempstead, New York: Hofstra Univ., 1965. (ERIC: ED 024 463.)

Taba, H., et al. Thinking in elementary school children. Cooperative Research Project No. 1574. San Francisco State College, 1964.

Thompson, G. R., & Bowers, N. C. Fourth grade achievement as related to creativity, intelligence, and teaching style. Paper presented at the meeting of the American Educational Research Association, Feb. 1968. For copies write to N. C. Bowers, School of Education, Northwestern Univ., Evanston, Ill.

Torrance, E. P., & Parent, E. Characteristics of mathematics teachers that affect students' learning. Cooperative Research Project No. 1020. Minneapolis: Minnesota School Mathematics and Science Center, Institute of Technology, Univ. of Minn., 1966. (ERIC: ED 010 378.)

Travers, R. M. W. Essentials of learning: An overview for students of education (2nd ed.), New York: Macmillan, 1967.

Travers, R. M. W., et al. Learning as a consequence of the learner's task involvement under different conditions of feedback. Journal of Educational Psychology, 1964, 55, 167-173.

Tyler, L. L. The concept of an ideal teacher-student relationship. Journal of Educational Research, 1964, 58, 112-117.

Ulrich, R., Stachnik, T., & Mabry, J. *Control of human behavior*. Glenview, Ill.: Scott, Foresman, 1966.

Vorreyer, D. F. An analysis of teacher classroom behavior and role. Unpublished doctoral dissertation, Univ. of Md., College Park, 1965.

Walberg, H. J. Predicting class learning: A multivariate approach to the class as a social system. *American Educational Research Journal*, 1969, 4, 529-543. (a)

Walberg, H. J. The social environment of learning as a mediator of classroom learning. *Journal of Educational Psychology*, 1969, 60, 443-449. (b)

Walberg, H. J. Curriculum evaluation: Problems and guidelines. *Teachers College Record*, 1970, 71, 557-571. (a)

Walberg, H. J. A model for research on instruction. *School Review*, 1970, 78, 185-200. (b)

Walberg, H. J., & Anderson, G. J. *Learning environment inventory*. Cambridge, Mass.: Harvard Project Physics, 1967.

Walberg, H. J., & Anderson, G. J. Classroom climate and individual learning. *Journal of Educational Psychology*, 1968, 59, 414-419.

Wallen, N. E. *Relationships between teacher characteristics and student behavior: Part three*. Cooperative Research Project No. SAE OE 5-10-181. Salt Lake City: Univ. of Utah, 1966. (ERIC: ED 010 390.)

Wallen, N. E., & Travers, R. M. W. Analysis and investigation of teaching methods. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally, 1963. Pp. 448-506.

Wallen, N. E., & Wodtke, K. H. *Relationships between teacher characteristics and student behaviors: Part One*. Cooperative Research Project No. 1217. Salt Lake City: Univ. of Utah, Dept. of Educational Psychology, 1963.

Wasik, B. H., Senn, K., Welch, R. H., & Cooper, B. R. *Behavior modification with culturally deprived school children: Two case studies*. Durham, N. Carolina: Duke Univ., The Education Improvement Program, 1968.

Welch, W. W. Curriculum evaluation. *Review of Educational Research*, 1970, 39, 429-445.

Welch, W. W., & Bridgham, R. G. Physics achievement gains as a function of teaching duration. *School Science and Mathematics*, May, 1968, 449-454.

Wilk, R. E., et al. *Student teacher activities and pupil responses.* Minneapolis: Univ. of Minn., Bureau of Educational Research, 1960.

Withall, J. The development of a technique for the measurement of social-emotional climate in classrooms. *Journal of Experimental Education*, 1949, 17, 347-361.

Wittrock, M. Set applied to student teaching. *Journal of Educational Psychology*, 1962, 53, 175-180.

Worthen, B. R. *Discovery and expository task presentation in elementary mathematics.* Journal of Educational Psychology Monograph Supplement, Vol. 59, No. 1, Part 2. Washington, D. C.: American Psychological Association, 1968. (a)

Worthen, B. R. A study of discovery and expository presentation: Implications for teaching. *Journal of Teacher Education*, 1968, 19, 223-242. (b)

Wright, C. J., & Nuthall, G. The relationships between teacher behavior and pupil achievement in three experimental elementary science lessons. *American Educational Research Journal*, 1970 (in press, November).

Yager, R. E. Teacher effects upon the outcomes of science instruction. *Journal of Research in Science Teaching*, 1966, 4, 236-242.

Zahorik, J. A. *Teacher verbal feedback and content development.* Paper presented at the meeting of the American Educational Research Association, Feb. 1969. Milwaukee: Univ. of Wis. (Mimeo.)